DART2010
4th International Workshop on Distributed Agent-based Retrieval Tools

Fourth International Workshop on

Distributed Agent-based Retrieval Tools

June 18, 2010
Webster University – Geneve, Switzerland

http://www.dart-project.org/dart2010/

Chairs: Vincenzo Pallotta, Alessandro Soro, and Eloisa Vargiu

DART2010
4th International Workshop on Distributed Agent-based Retrieval Tools

Proceedings of

DART2010

4th International Workshop on Distributed Agent-based Retrieval Tools

June 18, 2010

Webster University

Geneve, Switzerland

Workshop Chairs:

Vincenzo Pallotta

Alessandro Soro

Eloisa Vargiu

Table of Contents

KEYNOTE

CONTRIBUTIONS

Realizing Flexible and Pervasive Information Systems with HDS

Agostino Poggi and Federico Bergenti

Abstract HDS (Heterogeneous Distributed System) is a software framework that tries to simplify the realization of pervasive applications by merging the client-server and the peer-to-peer paradigms and by implementing all the interactions among the processes of a system through the exchange of typed messages and the use of composition filters for driving and dynamically adapting the behavior of the system. Typed messages and computational filters are the elements that mainly characterize such a software framework. In fact, typed messages can be considered an object-oriented implementation of the types of message defined by an agent communication language and so they are means that make HDS a suitable software framework both for the realization of multi-agent systems and for the reuse of multi-agent model and techniques in non-agent based systems. Composition filters, in HDS called message filters, drive and adapt the behavior of a system by acting on the exchange of messages. In fact, on the one hand, composition filters can constrain the exchange of messages (e.g., they can block the sending/reception of some messages to/from some processes), they can modify the flow of messages (e.g., they can redirect some messages to another destination) and they can manipulate messages (e.g., they can encrypt and descript messages). On the other hand, processes can dynamically add and remove some composition filters to adapt the behavior of a system to any hardware and software new configuration and to any new user requirement. The use of typed message and message filters together with the use of some multi-agent coordination techniques make HDS a suitable framework for realizing flexible and pervasive information management and retrieval systems able to adapt the system behavior both to the introduction of new kinds of information producer and consumers (e.g., new kinds of devices and new kind of information

Agostino Poggi
Dipartimento di Ingegneria dell'Informazione, University of Parma, Italy,
e-mail: poggi@ce.unipr.it

Federico Bergenti
Dipartimento di Matematica, University of Parma, Italy,
e-mail: Bergenti@CE.UniPR.IT

and new kinds of access rules (e.g., a producer offers information only to authorized consumers and through encrypted messages). The paper will introduce the HDS software framework and discussed its use for the realization of flexible and pervasive information management and retrieval systems.

Rethinking Search Engines in Social Network Vision

Manuela Angioni, Emanuela De Vita, Cristian Lai, Ivan Marcialis, Gavino Paddeu, Franco G. Tuveri

CRS4, Center of Advanced Studies, Research and Development in Sardinia, Parco Scientifico e Tecnologico, Ed. 1

09010 Pula (CA), Italy

{angioni, emy, clai, ciano, paddeu, tuveri}@crs4.it

Abstract In this paper we illustrate our vision about the evolution of search engines, dealing with some emerging questions related to the social role of the user on the Web and to the actual approach to access the information. In this scenario, is ever more evident the need to redefine the information paradigm bringing the information to the user and not more the user to the information, with search engines able to provide results without direct questions from users, anticipating their needs. A Web in service of the user, automatically informed by the system with suggested resources related with his life style and his common behaviour without the need to ask for them. This approach will be applied to a starting project named A Semantic Search Engine for a Business Network where the development of a business network creates a point of contact between the academic and the research world and the productive one by the introduction of Natural Language Processing, user profiling, automatic information classification according to users' personal schemas, contributing in such a way to redefine the vision of information and delineating processes of Human-Machine Interaction.

1 Introduction

The Web's evolution during the last few years shows that the advantages from the users' point of view are not so macroscopic. It is going more and more toward tools able to follow and assist the user in networking activities through the use of technologies related to natural languages, the classification of the information and the user profile (Marcialis and De Vita, 2008). In this scenario changes carried out by the great innovators in the field of information processing are emerging. Google is still the frontier of search engines, but there are several efforts in order to exceed its capabilities, such as Bing, which provides good results on search suggestions and allows natural language queries. Meta search engines try to reduce the time consumed on online search, allowing users to send queries simultaneously on more search engines and aggregating the results, such as BingandGoogle[1] or SortFix[2] , that searches Google, Yahoo and Twitter by means of a drag-and-drop interface that allows the user to describe a detailed and precise query.

In this paper we illustrate an overview and the ideas behind the starting project named *A Semantic Search Engine for a Business Network.* It involves the development of a business network able to create a point of contact between the academic and the research world in general and the productive one, with the aim of encouraging the cooperation and the sharing of ideas, of different point of views, information material or needs, and in order to support the productive world and the associated decision-making process. One of the project's objectives is to answer to the questions expressed in the following of the introduction.

Actually the online social networking is becoming more and more popular and several experiments in social network-based Web search have been performed in order to demonstrate the potential for using online social networks to enhance Internet search (Mislove, 2006).

The introduction of queries in natural language is a common element that is already prefiguring the advent of the Web 3.0. An example is Twine (Wissner and Spivack, 2009), able to improve the relevance of results by means of filters that try to reduce the noise due to less relevant answers. Other emerging tools are the computational knowledge engine Wolfram Alpha[3], able to answer queries by means of a vast repository of data organized with the help of sophisticated Natural Language Processing algorithms, or Aardvark (Horowits and Kamvar, 2010) that allows users, experts on certain topics, to answer to queries made by other users in a more efficient mechanism for online search.

Despite information is still the primary element, is ever more evident the need to redefine the information paradigm so that the net and the information become "really" user-centric by an inverse process that brings the information to the user and not more the user to information.

In our opinion, what each user needs is a specific private data strictly related to his point of view, his way to classify and manage the information, his network of contacts in the way everybody choose to live the Web, the net and the knowledge. So, new tools able to reduce or even to eliminate the search phase performed by the user are needed, but certainly commercial search engines, that make profit by the number of access to their pages, are not interested in produce them.

[1] http://www.bingandgoogle.com/

[2] http://www.sortfix.com/

[3] http://www.wolframalpha.com/

The passage from the unstructured to the structured information through the use of ontologies has not produced the expected innovation in search engines due to the lack of tagged resources.

The rethinking of search engines involves the emerging of some questions about the method of search through repeated queries and their successive refinement. Someone thinks that search engines should be considered “only a primitive form of decision support” (Spivack, 2010). So, the vision of a Web where search engines are able to provide results without direct questions from users, anticipating their needs, could be now plausible. A Web of user disposal, automatically informed by the system with suggested resources related with to his life style and his common behaviour without the need to ask for them. Such idea of Web and of search engines above described is applied to a project in starting phase and will converge in a system able to support and follow users in their activities. In particular the idea behind the project is the realization of a business network able to guarantee the match and the cooperation of academic and research world with the productive one in order to sustain related production and decisional processes.

The reminder of the paper is organized as follows: section 2 describes a use case. In section 3 are described the aims of the project in a general way; section 4 analyzes the search engine; section 5 shows the set of modules; section 6 explains the idea of business network and finally are described the conclusions.

2 Use Case

Companies and researchers meet on the Web using the usual navigation tools.

The following use case describes the effort to make compatible the needs of researchers and companies, making easier the meeting of their common goals and favouring the transfer between the research knowledge and the companies needs that look for innovative ideas to apply to their own business model.

The user browses the Web with Mozilla Firefox, provided with a proper extension. The viewed pages are modified and displayed according to the user profile and his preferences. An intelligent agent observes the user browsing in real time, pointing out all the information that better meet the definition of his profile as interesting.

The user, after the registration, is identified by his account and access to his homepage and to the social network notice board. The registration enables the system to access to the notice board for the analysis of contents. The system analyzes the user contacts list in order to generate a network of users unrelated to the profiles similarity. During his browsing activity the user highlights, annotates, inserts tags and classifies the interesting portions of the visited Web pages.

Moreover, the system shows information related to the current page. Information could include profiles of people with similar interests, companies and projects profiles correlated to the user profile, annotations of parts of pages by similar users and friends and besides, similar pages.

Related to networks of expertise generated by users, the system is able to propose and present, automatically and in real time, the matches between demand and supply of those intangibles (interests/ expertise/ know-how) that all companies would like to sell.

The result of the activity converges in all the networks that intends to integrate in the system (Facebook, Yammer, LinkedIn or Xing) and where the user is already connected to.

The user explicitly provides the system with useful information for refining the profile in order to put in evidence his interest in the received information.

The user can access his profile to verify if it is really consistent or if a distorted image of his interests is emerging. The system is able to state the reason why a specific correspondence has been proposed.

3 Goals of the Project

With social networks, blogs, RSS and new features in search engines are all news in the ICT context if compared with some years ago. This changed the way to access the information. The trend, hopefully, is the definition of new tools developed in order to follow the user in his activity and support him with the automatic generation and delivery of contents without his explicit request and according to his interests.

The automatic categorization of information through a predefined taxonomy, organized in a hierarchical category system, is often a restrictive and forced path. The same resource could be classified in different ways from different people and the same user could place the same page under different categories according to the reading context or to the content of interest. The classification of a document is, as well, depending on the personal culture, experience and context of life. Moreover, documents are often achieved using heterogeneous contents, talk about several topics and are obviously related to several categories.

Otherwise, with the Web 2.0, folksonomy, social tagging and social bookmarking place the user as starting point in a categorization work where each user labels resources. This step moves from a hierarchical logic to a more simpler way where all tags are at the same level.

Moving from the user management of information to an automatic one, a classification system should be able to categorize information according to user preferences and to relate his classification to a common set of categories based on a predefined taxonomy. By means of a such categorization tool, each user manages in a personal way his bookmarks, accedes to a quantity of Web sites, about scientific, news, entertainment or other topics, selecting, choosing and categorizing through the system. The system has to be able to manage a flow of data coming from a big set of predefined channels and updatable depending on the user preferences. Channels should be social networks, blogs, RSS services, news services, Web sites and also search engines, selected by the user.

To categorize information from these channels and to deliver contents that meet user preferences by means of a match algorithm based on the user profile and the document classification is a crucial point. The user can see categories associated to each resource labeled and ordered according to his schema.

4 A Semantic Search Engine for a Business Network

The semantic search engine is intended as a support tool for users, an active assistant able to give in *real time* references for the use of the information, reporting as more interesting the information that might match with the personal interest specified in the user profile.

There are two kinds of users: *companies* and the *generic user*, including employees, researchers, professionals, and people having specific interests and skills, according to the resources associated with them and emerging by their daily activities, that the system is able to track.

The business network is a point of contact between the academic and research world in general and the productive one and defines a communication level between users belonging to a community. The business network facilitates the sharing of knowledge, ability, expertise, skills, interests and resources between users belonging to the community that need or are interested in specific topics. In fact, it is not always easy to rise these feature, especially the immaterial expertise. But even publications or ongoing or past projects in which someone is involved, are often dispersed between public databases, or can be found only in the intranet of each company, or sometimes exists only in the head of someone, and it is not easy to explicit them. All the members of the community are linked together by the net of their skills: they are both depository of expertise in the service of users who need it: on the other hand they can need skills (papers, suggestions, projects, contacts) that other members can make available. This can be achieved with the development of an application, running on the computer of the user, that filters his activities and modifies his status, walls and links of the social network that the user subscribed, according to his permissions. Simultaneously records the activities on the user database.

The aim is to encourage the cooperation and the sharing of ideas, of different point of views, information material or needs, and to support the productive world and decision-making connected with it.

5 The Client Application

Users are organized as a community, configured according to their activities, through the management and by reporting organized content, information dynamically updated and personalized according to the specific user profile. The system will provide access to sources of shared documentation, to monitoring data, to support tools for sharing information between users, to networks of contacts explicitly specified in the community. The application shares this data with the other users that subscribed the community so that each user, according to the settings and the permissions, should know which resources have been visited, from whom and when. The application communicates these information to a plug-in installed on the user's browser that alerts the user and updates the visualization of information according to his preferences.

The system manages the user profile in order to control how the user preferences evolve during sessions of work. Information is monitored at time interval and new sessions can modify user preferences.

The system starts with a predefined user profile and evolves subsequently, using text categorization tools in order to categorize resources that are actually read, saved, commented. Only in these cases the system will modify the user criterion of classification for subsequently analysis.

The system follows step by step the evolution of user interests and suggests him, through the analysis of his profile, topics of interest, documents, contacts, etc, according to his interests. Moreover, the system is able to associate user profiles to companies or project profiles, automatically generating in real-time networks of expertise based on several configurable parameters and requirements.

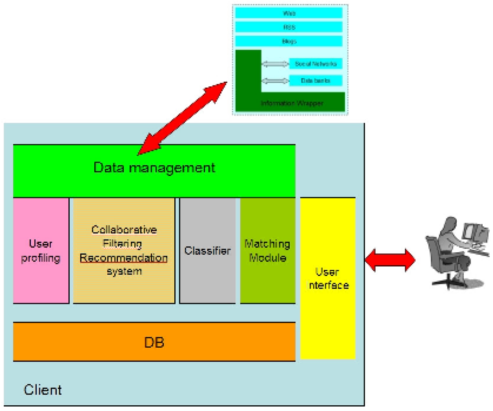


Fig. 1. The client application

Figure 1 shows a general description of the client application and the flow of the data coming from several distributed sources, such as social networks, blogs, RSS pages, visited Web pages, etc. There are four main modules: the User Profiling Module, the Collaborative Filtering and Recommendation System Module, the Classifier and the Matching Module, each responsible of the functionalities described below.

The level of communication between the modules and the distributed information is regulated by a layer that receives the data coming from the sources, and after an analysis and an opportune elaboration, is able to deliver to each module the portion of information that they are able to manage. Each module performs his activity, sometimes collaborating with other modules, and the result of the process is saved on a database. The interface allows queries in natural language and presents results according to the user profile and preferences. The system will be able to retrieve information from several textual and multimedia sources, and from Web services, even if conditionally.

Figure 2 shows in a summary way the data sources and a module named Information Wrapper that uniforms data coming from data banks (DBLP, ACM DL sites or institutional databases) and, under particular conditions, from social networks.

Some sources such as news services, social networks, blogs, RSS feeds, will be selected by the user or they will be automatically proposed by the system, by means of the preferences expressed by default or defined by the user profile and by the interests identified by the viewed pages.

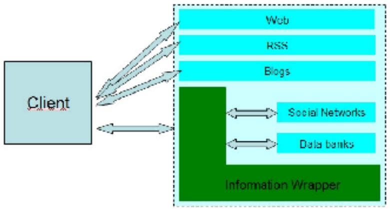


Fig. 2. The information wrapper

Other content will consist of personal and corporate profiles extracted from the HTML home pages, abstracts of scientific publications, bibliographies extracts from data banks.

More details of the modules involved in the system are described below.

5.1 Search Engine

The search engine module is contained in the Data Management Module, still under definition. The search engine indexes information coming from data sources and manages information related to the users, communities, companies, events, etc.

5.2 User Profiling

User profiling is a crucial process of the system because it has to define the user's interest, allowing the collaborative filtering and the recommendation tools to select and send information useful for the user itself. The module is able to classify and manage user information through the analysis of the resources he visited: the registration to rss resources, blogs, to social networks and the associated map of contacts, the collection of feedback, etc.
A profile for both users, companies and researchers, is defined creating in such a way a history depending on their activities and behaviour. So, the system will be able to identify user requirements and to predict its future behaviour and interests, in order to automatically propose resources useful to its activities without the need to search for them.
Data collected in this way are used by the system to find similarities, complementarities and links between companies and researchers, thus facilitating the match between supply and demand, particularly for intangibles such as interest, expertise, know-how.
The user should be able to access to its profile in order to check the reliability of the image that the system is bringing out, providing a positive or negative feedback to the matching proposed by the system.

5.3 Collaborative Filtering

During his activities, the user is supported by a module that helps him through two very important features: a collaborative filtering (De Vita et al., 2008) and a recommendation system. This module filters information by means of parameters based on the user preferences and his profile and gives advice to the user for news regarding communities and network activities that should be of interest. Advices are about:

- new activities
- users having similar interests
- companies having similar profile

- researchers having similar profile (based on their curriculum vitae)
- events of the network: workshop, conferences
- documents, papers, notes, projects, reviews classified that match users interests.
- announcements of competition, calls, etc

By means of the indications given by the user to the system it is possible to refine the profile.

5.4 Data Categorization

The system, with the user profile module, compares user profiles to company profiles through data categorization. It matches similar profiles, compares curricula of the user with request coming from companies, filters news and contents coming from the search engine working on the semantic of texts.
The classifier is based on a hierarchy of categories proposed by WordNet Domains (Magnini et al., 2002). These categories are the set of starting used by the system for the text categorization of resources.
The classifier performs a semantic disambiguation through the identification of relation between terms in order to identify composed terms, word sense disambiguation, name entities, geographic location.
The main phases are:

- Parsing of the text of resources (Web pages, documents, notes, etc)
- Analysis and syntactic disambiguation (Sleator and Temperley, 1993) (Liu, 2004)
- Semantic disambiguation and identification of real senses of words in sentences by means of a density function (Addis et al., 2009)
- Identification of name entities, geographic locations
- Classification of the textual resource by categories and values (Angioni et al., 2008a)
- Identification of semantic relations between concepts (Angioni et al., 2008b)

5.5 Matching module

The module is responsible to perform the matching between the information coming from the several data sources and by the users' profile, identifying those of real interest for each user.
It is able to organize data coming from users and companies profile, managing the textual resources, such as notes, papers, comments, profile data, previously analyzed by the classifier and aggregate the information.

Finally it sends notifications to users and the information as elaborated by the specific algorithm of matching.

Conclusion

The Web is changing and the way to access the information and the contents themselves are evolving too. Social networks, blogs, rss and new users' supports based on NLP are defining new evolutionary scenarios and creating new expectations for the Web. In this paper we illustrated a starting project named *A Semantic Search Engine for a Business Network* that defines a scenario where the above tools will converge in a system that, in our intention, will implement the use case described as a step of the vision described in the paper. The approach described aims at the development of the business network between the academic and the research world and the productive one, allowing a point of contact between users putting in evidence theirs skills and expertises.
The project aims both at implement the features described and at define and implement the described scenario. A validation to support the value of the expressed ideas will be one of the goal of the above mentioned project, where experimental results will be product.

References

Addis, A., Angioni, M., Armano, G., Demontis, R., Tuveri, F., Vargiu, E., 2008. A Novel Semantic Approach to Create Document Collections. In Antonio Palma dos Reis, editor, Proceedings Of Intelligent Systems And Agents Pages 53-60, 2008. IADIS Press. Selected for the best paper award.

Angioni, M., Demontis, R., Tuveri, F., 2008a. A Semantic Approach for Resource Cataloguing and Query Resolution. Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools, 5: 62-66.

Angioni, M., Demontis R., Deriu, M., Tuveri, F., 2008b. SemanticNet: a WordNet-based Tool for the Navigation of Semantic Information. In A.Tanacs, D.Csendes, V.Vincze, C.Fellbaum, and P.Vossen, editors, Proceedings Of GWC. University of Szeged.

De Vita, E., Deriu, M., Marcialis, I., Paddeu, G., 2008. Personalization and Collaborative Filtering for Information Retrieval on the Web. Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools, 5(-): 51-56.

Marcialis, I., De Vita, E., 2008. SEARCHY: An Agent to Personalize Search Results. A. Mellouk, editor, Third International Conference On Internet And Web Applications And Services. Volume -. Pages 512-517. IARIA. Institute of Electrical and Electronics Engineers (IEEE). Authorized distributor of all IEEE proceedings.

Horowits, D., Kamvar, S., 2010. The Anatomy of a Large-Scale Social Search Engine. Submitted to WWW2010, Raleigh, NC, USA.

Liu, H., 2004. MontyLingua: An end-to-end natural language processor with common sense, viewed 30 March 2010, <http://web.media.mit.edu/~hugo/montylingua>.

Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., 2002. The Role of Domain Information in Word Sense Disambiguation. Natural Language Engineering, special issue on Word Sense Disambiguation, 8(4), pp. 359 373, Cambridge University Press.

Mislove, A., Gummadi, K., Druschel, P., 2006. Exploiting Social network for Internet Search. In Proceedings of the 5th Workshop on Hot Topics in Networks, Irvine, CA.

Sleator, D.D., Temperley, D., 1993. Parsing English with a Link Grammar. in Third International Workshop on Parsing Technologies.

Spivack, N., 2010. Eliminating the Need for Search-Help Engines, viewed 30 March 2010, <http://www.novaspivack.com/uncategorized/eliminating-the-need-to-search>

Wissner, J., Spivack, N., 2009. Case Study: Twine. In W3C, Semantic Web Use Cases and Case Studies, viewed 30 March 2010, <http://www.w3.org/2001/sw/sweo/public/UseCases/Twine>

A Collaborative Web Application for Supporting Researchers in the Task of Generating Protein Datasets

Giuliano Armano and Andrea Manconi

Abstract The huge difference between known sequences and known tertiary structures has forested the development of automated methods and systems for protein analysis. When these systems are learned using machine learning techniques, the capability of training them with suitable data becomes of paramount importance. From this perspective, the search for (and the generation of) specialized datasets that meet specific requirements are prominent activities for researchers. To help researchers in these activities we developed ProDaMa-C, a web application aimed at *i*) generating specialized protein structure datasets and *ii*) favoring the collaboration among researchers. ProDaMa-C provides a collaborative environment where researchers with similar interests can meet and collaborate to generate new datasets. Datasets are generated selecting proteins through user defined pipelines of methods/operators. Of course each pipeline can also be used as starting point for building further pipelines able to enforce additional selection criteria. Freely available at the URL http://iasc.diee.unica.it/prodamac, ProDaMa-C has shown to be a useful tool for researchers involved in the task of generating specialized protein structure datasets.

1 Introduction

As the genome projects worldwide progress, the difference between known sequences and known tertiary structures increases exponentially. This huge discrepancy has fostered the development of automated methods and systems for pro-

Giuliano Armano
Dept. of Electrical and Electronic Engineering, University of Cagliari, Italy, e-mail: armano@diee.unica.it

Andrea Manconi
Dept. of Electrical and Electronic Engineering, University of Cagliari, Italy, e-mail: manconi@diee.unica.it

tein analysis. Systems to predict protein secondary structure (e.g. [1], [2]), transmembrane regions (e.g. [3]), and beta-turns (e.g. [4], [5]) are widely used. When these systems are generated using machine learning techniques, the capability of training them with suitable data becomes of paramount importance. From this perspective, the search for and/or the generation of specialized datasets that meet specific requirements are prominent activities for researchers. Different protein datasets have been proposed in the literature. However, these datasets are designed to investigate specific problems and may not be in accordance with the needs of researchers, or may not fit the specific nature of the problem. Owing to these limitations, researchers are often involved in the task of generating protein datasets, this task involves the problems of *i*) searching for, retrieving, and combining protein data from relevant specialized databases, *ii*) preprocessing and analyzing these data with suitable tools, and *iii*) overcoming the limitations associated with the migration of data and with the methods available for managing them. In our opinion, the social and collaborative nature of the Web 2.0, which encourages data integration as well as data sharing and reuse, is expected to provide a significant contribution to overcome the problems previously mentioned. In this perspective, we devised and developed ProDaMa-C (*Pro*tein *Da*taset *Ma*nagement - *C*ollaborative), a collaborative web application aimed at generating and sharing specialized protein structure datasets. ProDaMa-C is available for non-commercial use at http://iasc.diee.unica.it/prodamac/.

2 Methods

ProDaMa-C is a web application mainly aimed at making available a database of protein data, providing a repository of specialized bioinformatics tools, and supporting the collaboration among researchers involved in the task of generating specialized datasets. It has been developed using ProDaMa [6], a library of Python APIs devised to provide full support for generating protein structure datasets.

2.1 Protein Data

ProDaMa-C relies on a local database entrusted with storing information about protein data retrieved from a set of selected remote bioinformatics sources. In particular ProDaMa-C contains: *i*) protein data, downloaded from the Protein Data Bank (PDB) [7], *ii*) information about their classification from the CATH [8] and SCOP [9] databases, *iii*) information about membrane protein topologies from the MPTopo database [10], and iv) other information from the PDBFINDER database [11]. The local database has been pre-loaded with the proteins from PDB, as well as with a number of commonly used biological datasets (in particular RS126 [12], PDBSELECT25 [13], the PDB clusters of structures based on 50%, 70%, 90% and 95%

sequence identity, and the datasets of sequence structures used by WHAT IF [14] based on sequence identity, resolution and R-factor). All information stored in the local database is periodically updated.

2.2 Generating Datasets

With ProDaMa-C new datasets can be generated and made available starting from the content of the local database or from any previously-generated dataset. In both cases, the information source flows through a pipeline of operators. Any existing pipeline can be used as starting point to generate new pipelines able to provide further selection criteria. To generate a pipeline, three groups of operators are available off-the-shelf:

- *Search methods*, typically applied to select proteins that satisfy homology and/or similarity constraints. In particular, FASTA [15] and PSI-BLAST [16] services, useful to perform search by sequence similarity, are available, as well as PISCES [17], aimed at performing search by sequence identity. Methods for CATH and SCOP protein similarity searching, as well as for transmembrane protein topology search, are also provided. Furthermore, proteins can be selected by imposing constraints on their quality –i.e., on the experimental method that has been used, on the X-ray resolution, as well as on their R-factor and free R-factor.
- *Filter operators*, aimed at selecting relevant proteins according to a unary or binary predicate. In particular, given the input dataset, it is possible to select proteins according to the following constraints: *i*) single / multiple chains, *ii*) sequence length (e.g. length ≤ 200), *iii*) protein structure (e.g. number and/or length of alpha-helices or beta-strands, number of transmembrane segments), *iv*) percent of identity (e.g. sequence identity $\leq 25\%$).
- *Set operators*, as the classical union, intersection, and difference.

2.3 Collaborating with Other Users

According to the Web 2.0 philosophy, ProDaMa-C has been devised to support the collaboration among researchers with similar interests, in particular throughout the creation of groups of work aimed at dealing with common projects. To this end, users can build and modify their profile, manage a personal knowledge repository, create new projects, and join projects created by other users. Each user has an associated personal dashboard to manage pipelines and related databases, projects (including membership in other projects), and documents uploaded to the system. Users with similar interests can also create groups to work on common projects. Each project has an associated knowledge repository, where researchers can share ideas, documents, datasets, as well as pipelines useful to generate datasets in accordance with their needs. To facilitate the collaboration among researchers, a user can characterize her/himself with suitable tags (free-text keywords). The same tagging strategy can also be used to categorize projects. In so doing, users can easily discover other users and/or relevant projects related to their research interests. It is worth pointing out that, while both registered and anonymous users can access all resources of ProDama-C, only registered users can update them.

3 Conclusions and Future Work

Protein sequence analysis is an important research area in bioinformatics, owing to the huge difference between the number of known sequences and known tertiary structures. This discrepancy has promoted the development of automated methods of analysis, the accuracy of these systems being related to the data used in the training phase. From this perspective, researchers are often involved in the task of searching for and/or generating specialized protein datasets. To help them in these activities, we developed ProDaMa-C, a collaborative web application designed to generate specialized dataset according to user-defined criteria, and to support the collaboration among researchers. As for future work, we are planning to embed in ProDaMa-C other specialized tools for protein sequence analysis and for dataset management. In particular, tools for splitting the generated datasets into training and test set, as well as for supporting *k*-fold cross validation, will be provided soon. Furthermore, we plan to improve the collaborative environment of ProDaMa-C by collecting from the community comments and feedbacks about shared datasets and pipelines.

4 Acknowledgements

This work has been supported by the Italian Ministry of Education – Investment funds for basic research, under the project ITALBIONET – Italian Network of Bioinformatics. A special thanks goes to Stefano Curatti, for his help in implementing the library.

References

1. Jones D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Mol. Biology 1999, 292:192–202.
2. Pollastri G., Przybylski D., Rost B., Baldi P.: Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002, 47:228–235.
3. Randall A., Cheng J., Sweredosk M., Baldi P.: TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins. Bioinformatics 2008,

24(4):513–520.

4. Shepherd A.J., Gorse D., Thornton J.M.: Prediction of the location and type of beta-turns in proteins using neural networks. Protein Science 1999, 8:1045–1055.
5. Kaur H., Raghava G.P.S.: Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Science 2003, 12:627–634.
6. Armano G., Manconi A.: ProDaMa: an open source Python library to generate protein structure datasets. BMC Res. Notes 2009, 2:202.
7. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.: The Protein Data Bank. The Protein Data Bank 2000, 28:235–242.
8. Cuff A.L., Sillitoe I., Lewis T., Redfern O.C., Garratt R., Thornton J., Orengo C.A.: The CATH classification revisitedarchitectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleid Acids Research 2009, 37:D310–D314.
9. Andreeva A., Howorth D., Chandonia J.M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G.: Data Growth and its Impact on the SCOP Database: new Developments. Nucleic Acids Res. 2008, 36:D419–D425.
10. Jayasinghe S., Hristova K., White S.H.: MPtopo: A database of membrane protein topology. Protein Science 2001, 10:455–458.
11. Hooft R.W.W., Sander C., Scharf M., Vriend G.: The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. Bioinformatics 1996, 12(6):525–529.
12. Rost, B., Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy, Journal of Mol. Biology, 232, 584-599.
13. Hobohm, U., Sander, C. (1994) Enlarged representative set of protein structures, Protein Sci., 3(3), 522-524.
14. Vriend, G. (1990) WHAT IF: A molecular modeling and drug design program, J. Mol. Graph., 8, 52-56.
15. Pearson W.R., Lipman D.J.: Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1998, 85(8):2444–2448.
16. Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W., Lipman D.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997, 25(17):2289–3402.
17. Wang G., Dunbrack R.L.: Jr. PISCES: a protein sequence culling server. Bioinformatics 2003, 19:1589–1591.

Sensor Mining for User Behavior Profiling in Intelligent Environments

A. Augello, M. Ortolani, G. Lo Re, and S. Gaglio

Abstract The proposed system exploits sensor mining methodologies to profile user behaviors patterns in an intelligent workplace. The work is based in the assumption that users' habit profiles are implicitly described by sensory data, which explicitly show the consequences of users' actions over the environment state. Sensor data are analyzed in order to infer relationships of interest between environmental variables and the user, detecting in this way behaviour profiles. The system is designed for a workplace equipped in the context of Sensor9k, a project carried out at the Department of Computer Science of Palermo University.

1 Introduction

Research in Ambient Intelligence (AmI) focuses specifically on users and on how they relate to the surrounding environment; namely AmI systems attempt to sense the users' state, anticipate their needs and adapt the environment to their preferences [1]. Such systems usually rely on specific hardware in order to gather information about the environment state; such data may for instance be collected via pervasively deployed wireless sensor nodes [2], i.e. small devices equipped with sensors, a processor and a transceiver unit.

The idea presented here aims to investigate how such collected data might be profitably used to identify and profile user habits. In this context, data mining techniques allow for an intelligent analysis of environmental data in order to detect behavior patterns and classify them into profiles. Careful processing of sensory data may be used to infer descriptive models showing the relationships of interest between environmental variables and the user, while predictive models may proved

Agnese Augello, Marco Ortolani, Giuseppe Lo Re and Salvatore Gaglio
DINFO Dept. of Computer Engineering. University of Palermo, Viale delle Scienze, ed. 6 – Palermo, Italy, e-mail: (augello, ortolani)@dinfo.unipa.it, (lore, gaglio)@unipa.it

reliable inference on future behavior of users populating the considered environment [3].

User profiling applications in AmI have targeted environment personalization as for instance in [4], where a reinforcement learning algorithm is used to learn preferred music and lighting settings, adaptable to preferences changes. User profiling can also be used to detect significant changes in resident's behavior preserving their safety [10][12]. Other applications regard personalization of building energy and comfort management systems [11]. In [5], data collected by wireless sensor are used to create profiles of the inhabitants, and a prediction algorithm allows the automatic setting of system parameters in order to optimize energy consumption.

In this work sensor mining methodologies are exploited to profile user behaviors in an intelligent workplace. The workplace has been equipped in the context of Sensor9k, a project carried out at our Department [6]. Our work is based in the assumption that users' habit profiles are implicitly described by sensory data, which explicitly show the consequences of users' actions over the environment state. The system analyzes temporal data collected by the sensors located in the workplace rooms, and through a data mining process tries to detect changes which can be considered as consequences of user actions. Moreover the sensory data and the recognized events are arranged in appropriate models in order to highlight the existence of relationships among environmental data or events and the users' presence in the office room. The emerging behavioral patterns may finally be grouped based on their relative similarities by means of a clustering process in order to draw users profiles.

2 System Architecture

The proposed system aims to learn users' behavior profiles in the context of a smart workplace, such as that of the Sensor9k project [6]. Office rooms have been equipped with sensor nodes monitoring indoor and outdoor physical quantities such as relative humidity, temperature, and light exposure; additionally, RFId sensors allow for detecting the employees' presence in the workplace through the use of personal badges.

The overall system architecture, shown in Figure 1, has been designed according to a modular approach. A *preprocessing* module is used to improve the quality of sensory data while an *action detection* module analyzes data trends to infer changes which can be ascribed to human actions. The information extracted by sensors, and the recognized actions are arranged in appropriate models also keeping into account parameters of interest, such as temporal information or any particular environmental conditions. A *correlation* module is devoted to find relationships among the information described in the models; finally a *clustering* module allows to classify the patterns extracted by the correlation module.

The following sections will outline the most relevant features of each of the mentioned components.

2.1 Data Preprocessing

The data collected by the sensors are often affected by errors, due to imprecise measurements or to environmental noise. This module is devoted to the detection and removal of invalid values in raw sensory data and to possibly estimating missing data. Initial filtering can be performed assuming that there exist some admissible ranges for the values of the observed variables, and removing all values out of that range. Moreover, spacial and temporal redundancy can be exploited to detect anomalies in data, or to replace missing data. In particular temporal redundancy consists in the correlation between consecutive observations read from a sensor, while spatial correlation regards readings from neighboring sensors at a given time [3]. Temporal correlation can be exploited for the estimation of missing data by means of a linear interpolation between preceding and subsequent observations. Spatial correlation, on the other hand, can be profitably employed in order to detect outliers, and to replace them with the combination of neighbor sensors readings.

Subsequently, the set of observations is to undergo a dimensionality reduction process, by means of PCA [13]. By a linear transformation of the original variables, we extract the so-called principal components, by arranging them according to decreasing variance values. We can then easily remove information which contribute less to the variance of data, and are thus less relevant. The information loss due to the dimensionality reduction can be associated to noise in data. As an example, let $\mathbf{T} = [\mathbf{T}_1, \ldots \mathbf{T}_n]$ be a $m \times n$ matrix composed of a set of column vectors, each one representing the set of observations regarding the temperature measured by n sensors in an office room. After performing PCA we obtain a $m \times q$ matrix, with $q \leq n$, $\mathbf{T}' = [\mathbf{T}'_1, \ldots \mathbf{T}'_q]$.

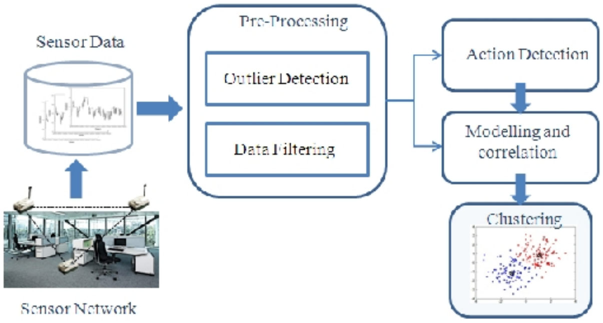


Fig. 1 System Architecture

2.2 Action Detection

This module has the aim to perform a deeper analysis of the observed variables trends and recognize those events that can be ascribed to human intervention. We assume here that sudden changes in observed values can be consequence of users' actions, such as turning on/off the light or changing the settings for the temperature and humidity control systems. As an example, Figure 2 shows the daily trend of the light exposure as measured by one of the sensor in an office room, and the corresponding relevant variations computed as derivative of the light function, which presumably correspond to actions on part of some user.

Probabilistic models, based on dynamic Bayesian networks, are then used to estimate which of those events may be in fact associated to human actions. As an example, Figure 3 shows the Bayesian network used to estimate the user actions controlling ambient light. The actions are modeled as states of the `UserAction t` variable, which depends on the state of the light (`LightTrend t` variable) and by information on the users presence (`UserPresence` variable) in workplace. The state of `UserAction t` variable and the state of the outdoor light `OutdoorLight t+1` influence in turns the state of the light at the next istant (`LightTrend t+1` variable).

2.3 Correlation

The correlation module is devoted to identifying potential relationships among the extracted information. Matricial models are used to represent the values recorded

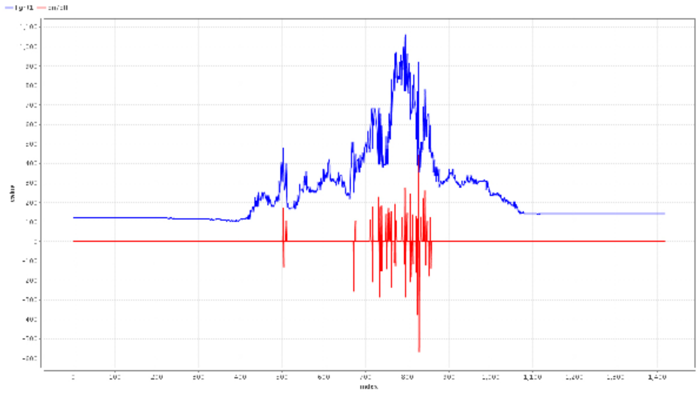


Fig. 2 **Light Relevant Events: red and blue functions represents respectively the trend of Light variable and the recognized events**

by different sensors for what concerns a given physical variable during a specific period (such as for instance an entire working day) and to represent the occurrences of events such as user actions in specific instants. In particular we can represent variables and events observations in a matrix $\mathbf{X}(m \times n)$, where a row $\mathbf{X}_i$ represents an observations at a specific time i and a specific column $\mathbf{X}_j$ represents the entire sample of observations of the j-th variable in the considered period.

In our specific case, matrix $\mathbf{X}$ is given by:

$$\mathbf{X} = \left[\mathbf{U}_1, \ldots \mathbf{U}_d, \mathbf{T}_1, \ldots \mathbf{T}_f, \mathbf{L}_1 \ldots \mathbf{L}_l\right]$$

composed of a set of vectors, each one represents the set of observations of a specific variable. In particular the set $\mathbf{U} = \{\mathbf{U}_j\}_{j=1\ldots d}$ represents observations about the presence of d users in the considered period, while sets $\mathbf{T} = \{\mathbf{T}_j\}_{j=1\ldots f}$ and $L = \{\mathbf{L}_j\}_{j=1\ldots l}$ represent observations about temperature and light exposure, respectively, related to the f and l variables obtained after the application of PCA on temperature and light matrices as described in Section 2.1.

The correlation matrix $\mathbf{R}(n \times n)$ is then computed in order to highlight the relationships among the variables. The i,j-th element of $\mathbf{R}$ is given by the correlation coefficient r_{ij} between the i-th and the j-th variable, as given by:

$$r_{ij} = Corr(\mathbf{X}_i, \mathbf{X}_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

where σ_{ij} is the covariance between $\mathbf{X}_i$ and $\mathbf{X}_j$ and σ_i and σ_j are respectively the standard deviation of $\mathbf{X}_i$ and $\mathbf{X}_j$.

In this way is possible extract correlation patterns between the observations related to the presence of a the users in office rooms and values representative of specific environment variables. In this way it is possible obtain a characterization of users with respect to values of the observed variables in a specific period.

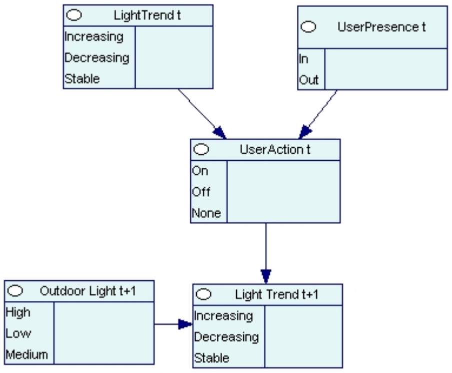


Fig. 3 **A Probabilist model to estimate user actions controlling ambient light**

2.4 Clustering

The clustering module allows to classify the pattern extracted by the correlation module. The clustering leads to the subdivision of users' behavior patterns into a set of profiles based on their similarities. In this way we can group users with similar preferences about variables setting, or users performing the same actions in similar environment conditions. In particular the K-means [14] algorithm can be used to classify data from the submatrices extracted from the correlation matrix **R**. In the current implementation, this module is in a preliminary phase. The aim is to determine an appropriate number of clusters and to choose an adequate distance function to evaluate distances between data points and cluster centers.

3 Conclusion and future works

In this paper we presented a sensor mining system aimed to profiling occupant behaviours. Environmental variables are monitored by a sensor network, and a set of modules allow to extract useful information regarding user actions and habits. The system is still under development, therefore future works will regard a deeper refinement of the system modules and the evaluation of the proposed approach for the case of study of Sensor9k.

References

1. A. Vasilakos and W. Pedrycz. Ambient Intelligence, Wireless, Networking, Ubiquitous Computing. Artech House Press, MA, USA (2006).
2. Akyildiz, IF and Su, W. and Sankarasubramaniam, Y. and Cayirci, E.: A survey on sensor networks. IEEE communications magazine. Volume 40, Issue 8, 2002, Pages 102–114
3. Shaomin Wu, Derek Clements-Croome, Understanding the indoor environment through mining sensory data–A case study, Energy and Buildings, Volume 39, Issue 11, November 2007, Pages 1183-1191, ISSN 0378-7788, DOI: 10.1016/j.enbuild.2006.07.011.
4. A. Khalili, C. Wu and H. Aghajan. Autonomous Learning of User's Preference of Music and Light Services in Smart Home Applications. Behavior Monitoring and Interpretation Workshop at German AI Conf, Sept 2009.
5. Barbato, L. Borsani, A. Capone, S. Melzi. Home Energy Saving through a User Profiling System based on Wireless Sensors. ACM Buildsys 2009 (in conjunction with SenSys 2009), Berkeley, CA, Nov. 3, 2009.
6. Alessandra De Paola, Alfonso Farruggia, Salvatore Gaglio, Giuseppe Lo Re, Marco Ortolani: Exploiting the Human Factor in a WSN-Based System for Ambient Intelligence. CISIS 2009: 748-753
7. Alessandra De Paola, Salvatore Gaglio, Giuseppe Lo Re, Marco Ortolani: Human-ambient interaction through wireless sensor networks. Proceedings of the 2nd IEEE conference on Human System Interactions, Pages 61–64, 2009.
8. M.J. Akhlaghinia, A. Lotfi, C. Langensiepen, and N. Sherkat. Occupant Behaviour Prediction in Ambient Intelligence Computing Environment. Special Issue on Uncertainty-based Technologies for Ambient Intelligence Systems. Volume 2 Number 2, May 2008.

9. T. Fawcett and F. J. Provost. Combining data mining and machine learning for effective user profiling. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), pages 8–13, 1996.
10. Diane J. Cook, Juan C. Augusto, Vikramaditya R. Jakkula. Ambient intelligence: Technologies, applications, and opportunities. Pervasive and Mobile Computing, Volume 5, Issue 4, August 2009, Pages 277-298, ISSN 1574-1192, DOI: 10.1016/j.pmcj.2009.04.001.
11. Dong, B. and Andrew, B. Sensor-based Occupancy Behavioral Pattern Recognition for Energy and Comfort Management in Intelligent Buildings. Proceedings of Building Simulation '2009, an IBPSA Conference, Glasgow, U.K.
12. M.J. Akhlaghinia, A. Lotfi, C. Langensiepen, and N. Sherkat. Occupant Behaviour Prediction in Ambient Intelligence Computing Environment. Special Issue on Uncertainty-based Technologies for Ambient Intelligence Systems. Volume 2 Number 2, May 2008.
13. Jolliffe, I. T. Principal Component Analysis. Springer-Verlag. pp. 487. (1986) doi:10.1007/b98835. ISBN 978-0-387-95442-4.
14. J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, (1967) 1:281-29

Content novelty and group recommendation

Ludovico Boratto, Salvatore Carta, Michele Satta

Abstract Recommender systems usually propose items to single users. However in some contexts and domains it might be impossible to make recommendations for each user, because of limitations imposed by the system. In [2] a group recommender system able to detect intrinsic communities of users whose preferences are similar was proposed. However, there are types of recommended items (like movies) that should always be new, i.e., it wouldn't make sense to recommended an item if a great part of the group has already expressed a preference for it (e.g., a movie already seen by a lot of users of the same group). This paper will focus on how novelty of the recommended items affects the quality of the recommendations.

1 Introduction

Recommender systems aim to provide information items that are expected to interest a user [5].

There are contexts and domains where, however, classic recommender systems cannot be used. For example:

- in multiple access systems with limited transmission capacity like Satellite Systems, it might not be possible to create personalized program schedules;

Ludovico Boratto
Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72 - 09124 Cagliari, Italy e-mail: boratto@sc.unica.it

Salvatore Carta
Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72 - 09124 Cagliari, Italy e-mail: salvatore@unica.it

Michele Satta
Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72 - 09124 Cagliari, Italy e-mail: michele_satta@hotmail.com

- in a website's homepage, the limited space available is usually divided into sections that contain different types of content (e.g., sport, entertainment, etc) to create interest for everyone.

Group recommendations have already been studied from several perspectives. In [4] and [3] the state-of-the-art in group recommendation is presented and techniques developed for different domains like web/news pages, tourist attractions, music tracks, television programs and movies are described. Nearly all the existing approaches take for granted the presence of groups of users with similar opinions but, in reality, this information is usually not available. In [2] an algorithm to generate group recommendations, able to automatically detect groups of users whose preferences are similar, was proposed.

A group recommendation approach that recommends the same content previously evaluated by users would be useful for content that is always renewed and ever-changing, like news items or TV series episodes. Users preferences for such types of content can be used to recommend items of the same type (e.g., news about the same topic or new episodes of the same series).

On the contrary, when a system produces group recommendations for types of content like movies, a new issue arises: *novelty* of the recommended items. In fact, if an item was already evaluated by a great part of the group, the system should limit its recommendation: users who already considered the item would be bored to watch/read/listen to it often and it wouldn't be a real recommendation for them.

This paper will present a study that shows how novelty of the recommended content affects the quality of group recommendations. Recommending novel content creates a trade-off that involves an improvement in satisfaction of the users and a loss in the quality of the predicted ratings. Since groups of different sizes are automatically detected by the system we used, this study would allow a content provider to explore such a trade-off, considering also the level of personalization of the recommended content. To the best of our knowledge, this is the first study of this type.

The rest of the paper is organized in the following way: section 2 contains a description of the system used in this study; section 3 describes the experiments we conducted and outlines main results; section 4 will draw some conclusions.

2 Content novelty and group recommendation

Here we will briefly introduce the group recommendation algorithm used in this study. The algorithm works in four steps:

2.1 Users similarity evaluation

In order to create communities of users, the algorithm takes as input a *ratings matrix* and evaluates through a standard metric (cosine similarity) how similar the preferences of two users are. The result is a weighted network where nodes represent users and each weighted edge represents the similarity value of the users it connects. A post-processing technique is then introduced to remove noise from the network and reduce its complexity.

2.2 Communities detection

To identify intrinsic communities of users, a Community Detection algorithm proposed by [1] is applied to the users similarity network and partitions of different granularities are generated.

2.3 Ratings prediction for items rated by enough users of a group

A group's ratings are evaluated by calculating, for each item, the mean of the ratings expressed by the users of the group. In order to predict meaningful ratings, our algorithm calculates a rating only if an item was evaluated by a minimum percentage of users in the group. With this step it is not possible to predict a rating for each item, so another step has been created to predict the remaining ratings.

2.4 Ratings prediction for the remaining items

For some of the items, ratings could not be calculated by the previous step. In order to estimate such ratings, similarity between items is evaluated, and the rating of an item is predicted using a classic Item-Based Nearest Neighbor Collaborative Filtering Algorithm proposed in [6].

3 Experimentation

The main objective of our experiments is to measure how much novelty of the recommended content affects the quality of the group recommendation, considering different partitions of the users in groups. To make the study, we used the Movie-

Lens[1] 10M/100K dataset (composed of 10 million ratings, expressed by 69878 users for 10681 movies). We built a framework that extracts a subset of ratings from the dataset, predicts group recommendations through the algorithm described in 2 and measures the quality of the predictions.

3.1 Experimental methodology and setup

Around 20% of the ratings was extracted as a test set and the rest was used as a training set for the algorithm. For each partition of the users in groups, ratings were predicted and the quality of the predictions was evaluated through the Root Mean Squared Error (RMSE). The metric compares the test set with the predicted ratings: each rating r_i expressed by a user u for an item i is compared with the rating $\bar{r}_i$ predicted for the item i for the group in which user u is.

RMSE is a metric widely used to evaluate the quality of recommendations. To evaluate the RMSE values we obtained, we considered the range between recommendations made for a single user and recommendations made for a single group that contains all the users. Inside that range it is reasonable to compare the different partitions, considering that recommendations predicted for a single user are the best result that can be obtained (predictions are tailored to a user's preferences) and a broadcast recommendation for a single group with no novelty of the content is still acceptable.

In each experiment we evaluated the system performances considering different values of a *novelty* parameter, i.e., the minimum percentage of users in a group that didn't previously rate an item, in order for it to be recommended. For example, if *novelty* was set to 50% and an item was rated by 60% of the group, the predicted rating for that item would be discarded, since it wouldt be novel just for 40% of the group.

To evaluate how the performances of the group recommendation algorithm varied for different values of the *novelty* parameter, we compared them with the results obtained using a classic User-Based Nearest Neighbor Collaborative Filtering Algorithm proposed in [6], where recommendations are produced for each user.

3.2 Experimental results

Fig. 1 shows RMSE for different values of content novelty for a group, considering different partitions of the users in 1, 4 and 5 groups. Partitions in 4 and 5 groups were the output of Step 3 presented in 2. As briefly explained, we wanted to consider intrinsic communities of users, i.e., a real partition of the network without any constraint given to the algorithm. However, it would be interesting to have a sce-

[1] http://www.grouplens.org/

nario with different partitions, so one of the future developements of this algorithm will allow us to have more partitions of the network, by adding some constraints to the algorithm used to identify the groups.

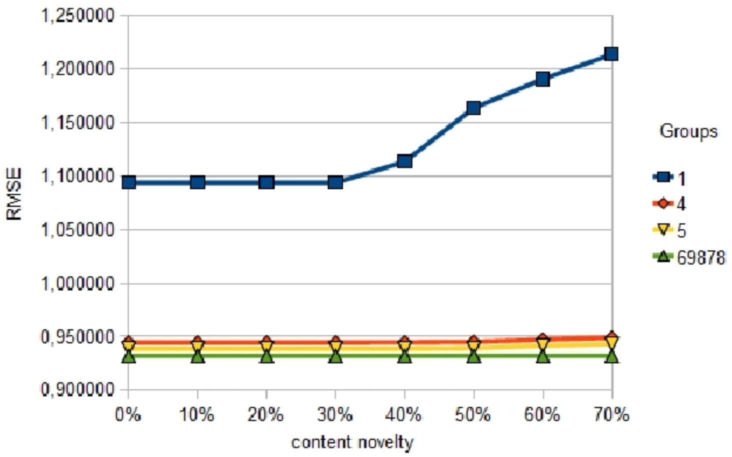


Fig. 1 RMSE for different values of novelty

The case in which there is a single group and the same content is recommended to all the users is the worst (compared to the approach in which recommendations are predicted for every user, worsening in RMSE values is between 17% and 30%). In order to highlight our algorithm's performances, the worsening percentage of RMSE for 4 and 5 groups compared to the optimal case in which users are not grouped (the green line in Fig. 1), is presented in Fig. 2. We can see that the partition in 5 groups (worsening percentage between 0,71% and 1,19%) performs slightly better than the partition in 4 groups (worsening percentage between 0,74% and 1,22%). We can notice that in all the studied partitions performances drastically worsen when *novelty* $> 40\%$. This is due to the large size of the groups for the partitions we considered. In this case the number of movies that were not evaluated by a large percentage of the users is very limited, so the movies available for the recommendation suddenly decrease, limiting the quality of the recommendation. With smaller groups, this novelty threshold (which now is around 40%) would be higher.

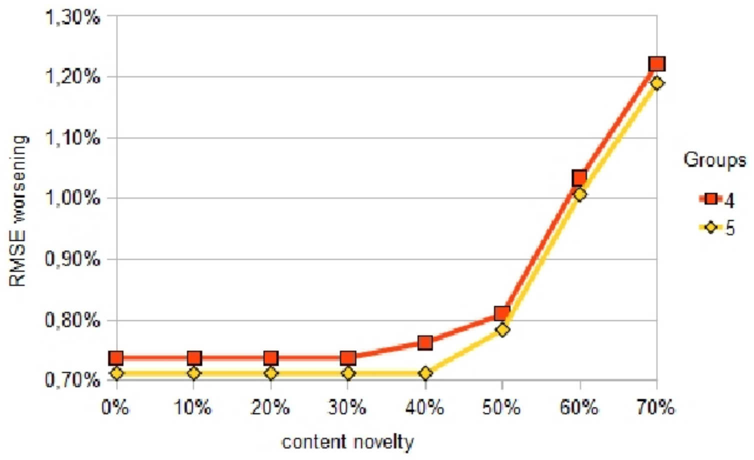


Fig. 2 Performances with partitions in 4 and 5 groups

4 Conclusions and future work

This paper presented a study to evaluate how novelty of the recommended content affects the quality of group recommendations. Experimental results show an interesting scenario with different partitions of users in groups, that would help a content provider configure the system, evaluating the trade-off between the different aspects.

Acknowledgment

This work is partially funded by Regione Sardegna under project CGM (Coarse Grain Recommendation) through Pacchetto Integrato di Agevolazione (PIA) 2008 "Industria Artigianato e Servizi".

The authors would like to thank Matteo Desogus for participating in the definition and implementation of the group recommendation algorithm.

References

1. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008+, October

2008.

2. Ludovico Boratto, Salvatore Carta, Alessandro Chessa, Maurizio Agelli, and M. Laura Clemente. Group recommendation with automatic identification of users communities. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 3:547–550, 2009.
3. Ludovico Boratto and Salvatore Carta. State-of-the-Art in Group Recommendation and New Approaches for Automatic Identification of Groups. In Soro, A., Vargiu, E., Armano, G., Paddeu, G., eds. *Information Retrieval and Mining in Distributed Environments* In press.
4. Anthony Jameson and Barry Smyth. Recommendation to groups. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer, 2007.
5. P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.
6. J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter 9, pages 291–324. 2007.
7. Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1, pages 210–217, 1995.

Opinion Mining and Sentiment Analysis Need Text Understanding

Rodolfo Delmonte[2] , Vincenzo Pallotta[1]

[1]Department of Computer Science
Webster University, Geneva
Route de Collex 15
CH-1293 Bellevue, Switzerland
pallotta@webster.ch

[2]Department of Language Science
Università "Ca Foscari"
30123 – Venezia, Italy
delmont@unive.it

ABSTRACT

We assume that in order to properly capture opinion and sentiment expressed in a text or dialog any system needs a deep text processing approach. In particular, we use Ontology matching and Concept search, based on SentiWordNet, but the system is required to spot fundamental issues as the following ones: presence of NEGATION at different levels of syntactic constituency; presence of LEXICALIZED NEGATION in the verb or in adverbs; presence of conditional, counterfactual subordinators; double negations with copulative verbs; presence of modals and other modality operators.

In order to cope with these linguistic elements we propose to build a Flat Logical Form (FLF) directly from a Dependency Structure representation augmented by indices and where anaphora resolution has operated pronoun-antecedent substitutions. We implemented these additions our the system called VENSES which has been used for semantic evaluation purposes in the challenge called RTE. The output of the system is an xml representation where each sentence of a text or dialog is a list of attribute-value pairs, one of which is POLARITY.

1 Introduction

We assume that in order to properly capture opinion and sentiment expressed in a text or dialog any system needs a deep text processing approach. In particular, the idea that the task may be solved by the use of Information Retrieval tools like Bag of Words Approaches (BOWs) is totally flawed. BOWs approaches are sometimes also camouflaged by a keyword based Ontology matching and Concept search, based on SentiWordNet, by simply stemming a text and using content words to match its entries and produce some result. Any search based on keywords and BOWs is fatally flawed by the impossibility to cope with such fundamental issues as the following ones:

- presence of NEGATION at different levels of syntactic constituency;
- presence of LEXICALIZED NEGATION in the verb or in adverbs;

- presence of conditional, counterfactual subordinators;
- double negations with copulative verbs;
- presence of modals and other modality operators.

In order to cope with these linguistic elements we propose to build a Flat Logical Form (FLF) directly from a Dependency Structure representation augmented by indices and where anaphora resolution has operated pronoun-antecedent substitutions. We implemented these additions our the system called VENSES which has been used for semantic evaluation purposes in the challenge called RTE. The output of the system is an xml representation where each sentence of a text or dialog is a list of attribute-value pairs, one of which is POLARITY. In order to produce this output, the system makes use of the FLF and a vector of semantic attributes associated to the verb at propositional level and memorized. Important notions required by the computation of opinion and sentiment are also the distinction of the semantic content of each proposition into two separate categories:

. OBJECTIVE vs SUBJECTIVE

This distinction is obtained by searching for FACTIVITY markers again at propositional level. In particular we take into account:
- tense, voice, mood at verb level
- modality operators like intensifiers and diminishers, but also modal verbs
- modifiers and attributes adjuncts at sentence level
- lexical type of the verb (in Levin's classes and also using WordNet classification)
- subject’s person (if 3rd or not).

2 The VENSES system

VENSES is a reduced version of GETARUNS (Delmonte, 2007 and 2008), a complete system for text understanding developed at the Laboratory of Computational Linguistics of the University of Venice. The backbone of VENSES is LFG theory in its original version (Bresnan, 1982 and 2000). The system produces different levels of analysis, from syntax to discourse. However, three of them contribute most to the opinion classification task: the lexico-semantic, the anaphora resolution and the deep semantic module.

2.1 The syntactic and lexico-semantic module

The system produces a c-structure representation by means of a cascade of augmented FSA, then it uses this output to map lexical information from a number of different lexica which however contain similar information related to verb/adjective and noun subcategorization. The mapping is done by splitting sentences into clauses which are main and subordinate clauses. Other clauses are computed in their embedded position and can be either complement or relative clauses.

The output of the system is what we call AHDS (Augmented Head Dependent Structure) which is a fully indexed logical form, with Grammatical Relations and Semantic Roles. The inventory of semantic roles we use is however very small – 35, even though it is partly overlapping the one proposed in the first FrameNet project. We prefer to use generic roles rather than specific Frame Elements (FEs) because sense disambiguation at this stage of computation may not be effective.

2.2 The anaphora resolution module

The AHDS structure is passed to and used by a full-fledged module for pronominal and anaphora resolution, which is in turn split into *two submodules*. The resolution procedure takes care only of third person pronouns of all kinds – reciprocals, reflexives, possessive and personal. Its mechanisms are quite complex, as described in (Delmonte et al., 2006). The *first submodule* basically treats all pronouns at sentence level – that is, taking into account their position – and if they are left free, they receive the annotation "external". If they are bound, they are associated to an antecedent's index; else they might also be interpreted as expletives, i.e. they receive a label that prevents the following submodule to consider them for further computation.

The *second submodule* receives as input the external pronouns, and tries to find an antecedent in the previous stretch of text or discourse. To do that, the systems computes a *topic hierarchy* that is built following suggestions by (Sidner and Grosz, 1986) and is used in a centering-like manner.

2.3 The semantic module

The output of the anaphora resolution module is used by the semantic module to substitute the pronoun's head with the antecedent's head. After this operation, the module produces Predicate-Argument Structures or PAS on the basis of previously produced Logical Form. PAS are produced for each clause and they separate obligatory from non-obligatory arguments, and these from adjuncts and modifiers. Some adjuncts, like spatiotemporal locations, are only bound at propositional level.

This module produces also a representation at propositional level, which for simplicity is just a simple vector of information containing 15 different slots, each one devoted to contain a different piece of semantic information. We encode the following items: modality, negation, focussing intensifiers/diminishers, manner adjuncts, diathesis, auxiliaries, clause dependency if any from a higher governing predicate – this is the case for infinitivals and gerundives – and eventually a subordinator if any.

2.4 The classification system

Differently from other systems, we use a three way classification for the attribute “attitute” which encodes polarity: POSITIVE, NEGATIVE and SUSPENSION. The latter is used when negation is present in the utterance but the overall attitude is not directly negative. More on this below. Bing(2004) uses a scale of three grades to indicate strength and distinguish cases of real NEGATIVE/POSITIVE polarity from one another. In the readme file associated to the datasets, Bing comments on this grading that “… note that the strength is quite subjective. You may want to ignore it, but only considering + and –“. It is a fact that annotation criteria are hard to establish, but then the outcome is always subjective in a sense. For instance in the example below taken from one of the datasets he makes available on his website, that we evaluated (see below), the score [-1] indicates a low negative polarity strength. We report at first Bing’s annotated example under A. and then our system’s output, under B.

A. viewfinder[-1]##the lens is visible in the viewfinder when the lens is set to the wide angle , but since i use the lcd most of the time , this is not really much of a bother to me .
B. id="44", predicate="be", topic="lens", attitude="suspension" factivity= "factive_statement"

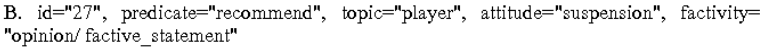
Here and elsewhere we annotated SUSPENSION, and the system correctly labels the example: this label indicates an attitude which is not strongly marked for either polarity value, and in some cases this may also be due to the presence of double negation. As a consequence, we also use SUSPENSION in the following example, where on the contrary, Bing annotates Positive with good confidence:

A. weight[+2]##at 8 ounces it is pretty light but not as light as the ipod .
B. id="46", predicate="be", topic="light", attitude="suspension" factivity= "factive_statement"

We assume that this case cannot be computed as a strong case for positive attitude. In many other cases, Bing does not produce any annotation which does not mean that the utterance is to be regarded as neutral, as for instance in the following cases,

A. ##if you have any doubts about this player , well do n't .
B. id="32",predicate="have" topic="player" attitude="suspension" factivity= "opinion_internal"

A. ##can 't complain and i recommend it over all the other players , just hope that remote will come out soon .

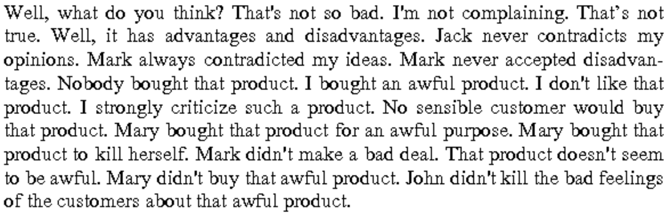


B. id="27", predicate="recommend", topic="player", attitude="suspension", factivity= "opinion/ factive_statement"

In many cases, however, it is hard to understand the reason why the annotation has not been made available as for instance in,

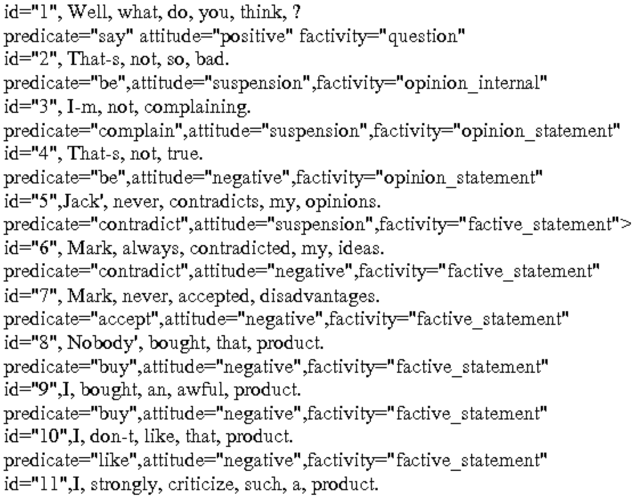
A. do not buy this player .
A. a piece of junk .
A. don't waste your money .
A. the g2 was better in this respect .

Here below we present a short made-up dialog which contains a certain number of complex negative cases to solve.



> Well, what do you think? That's not so bad. I'm not complaining. That's not true. Well, it has advantages and disadvantages. Jack never contradicts my opinions. Mark always contradicted my ideas. Mark never accepted disadvantages. Nobody bought that product. I bought an awful product. I don't like that product. I strongly criticize such a product. No sensible customer would buy that product. Mary bought that product for an awful purpose. Mary bought that product to kill herself. Mark didn't make a bad deal. That product doesn't seem to be awful. Mary didn't buy that awful product. John didn't kill the bad feelings of the customers about that awful product.

Here below is a schematic output from the xml representation of each sentence:



```
id="1", Well, what, do, you, think, ?
predicate="say" attitude="positive" factivity="question"
id="2", That-s, not, so, bad.
predicate="be",attitude="suspension",factivity="opinion_internal"
id="3", I-m, not, complaining.
predicate="complain",attitude="suspension",factivity="opinion_statement"
id="4", That-s, not, true.
predicate="be",attitude="negative",factivity="opinion_statement"
id="5",Jack', never, contradicts, my, opinions.
predicate="contradict",attitude="suspension",factivity="factive_statement">
id="6", Mark, always, contradicted, my, ideas.
predicate="contradict",attitude="negative",factivity="factive_statement"
id="7", Mark, never, accepted, disadvantages.
predicate="accept",attitude="negative",factivity="factive_statement"
id="8", Nobody', bought, that, product.
predicate="buy",attitude="negative",factivity="factive_statement"
id="9",I, bought, an, awful, product.
predicate="buy",attitude="negative",factivity="factive_statement"
id="10",I, don-t, like, that, product.
predicate="like",attitude="negative",factivity="factive_statement"
id="11",I, strongly, criticize, such, a, product.
```

```
predicate="criticize",attitude="negative",factivity="factive_statement"
id="12", No, sensible, customer, would, buy, that, product.
predicate="buy",attitude="negative",factivity="factive_statement"
id="13", Mary, bought, that, product, for, an, awful, purpose.
predicate="buy",attitude="negative",factivity="factive_statement"
id="14", Mary, bought, that, product, to, kill, herself.
predicate="buy",attitude="negative",factivity="factive_statement"
id="15", Mark, didn-t, make, a, bad, deal.
predicate="make",attitude="relevant",factivity="opinion_statement"
id="16", That, product, 'doesn-t_', seem, to, be, awful.
predicate="seem",attitude="suspension",factivity="opinion_internal"
 id="17",Mary,didn-t, buy, that, awful, product.
predicate="buy",attitude="negative",factivity="factive_statement"
id="18",John didn't kill the bad feelings of the customers about that awful product.
predicate="kill",attitude="suspension",factivity="factive_statement"
```

2.5 Distinguishing FACTIVITY from POLARITY

As can be easily noticed, the problem constituted by the presence of negation is not solvable by a simple one way decision – yes/no. In many cases the information about the attitude of the speaker is just not directly communicated and needs further specification. Sentence 16, for instance, is not a straightforward admission of disagreement; the same applies to sentence 18. We also regard sentences 2, 3, 5 to be cases of indirect judgement which however is not explicit enough to be assigned to a positive attitude. For this reason, we decided to introduce a marker, we call SUSPENSION which encodes all cases of indirect judgement, and other similar situations. Coming now to clear cases of NEGATIVE attitude, we register sentences like 4, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 17. However, not all these sentences are easily understood as such. In particular, 4 and 10 are simple cases of negation at main verb level and may be computed safely as cases of negative attitude. Sentences 6 and 11 are again cases of negative attitude but there is no explicit negation expressed: just negatively marked verb at lexical level. Examples 8 and 12 express negation at subject level: as can be gathered, this can only be evaluated as a real negative attitude only in case the main verb indicates positive actions. Apparently, these cases can also be contradicted by the same speaker, by using BUT and other adversative discourse markers (even though nobody likes it …; nobody likes it, but …). Examples 9, 13, and 14 introduce negative attributes at object and complement level. This is also computed by the system as a case of negative attitude. The system also computes as negative example 17, which is a case of double negation: in this sentence, negation is present both at verbal level and at complement level. This might be understood as positive attitude (if she did not do that then it is good…). However we assume that this is also intepretable as a report of something negative that might have happened.

Eventually we have cases of SUSPENSION envolving the presence of negation as example 18 shows.

An important subdivision of all semantic types involved regards FACTIVITY which as said above can constitute an important indicator of the speaker's attitude in uttering a given judgement. It is important to mention that sentences are in fact utterances which can be subdivided in at least two main types: they constitute OBJECTIVE or FACTIVE STATEMENTs reporting in this way some fact usually in third person subject; or they may constitute SUBJECTIVE or NON-FACTIVE OPINIONs expressed by the speaker him/herself in first person or reporting on somebody elses's opinion. Opinions are always subjective but may report an internal thought, a wish, a hope, or else a definite state, event, activity by the subject. In the former case, we OPINION_INTERNAL, to highlight the weight of subjective markers as in 2 and 16. In the latter case, we use OPINION_STATEMENT because it is either the case that the utterance refers acts or events of third persons, as in 15; or else, it reports the evaluation of the speaker as in 3 and 4.

Other markers are QUESTION which can still be computed as either positive or negative; and RELEVANT, implying some indirect judgement as shown by 16 and double negation and reinforcing on SUSPENSION.

Now consider a really hard utterance to evaluate:

*Positive-1 dvd - so far the dvd works so i hope it doesn't break down like the reviews i 've read .

This has been rightly annotated as Positive by Bing however for a system to compute the dependence of the NEGATED sentence from HOPE one needs a logical form and all the appropriate indices, to capture it.

2.6 The remaining semantic markers: CONDITIONAL and COMPARATIVES

Eventually, there are important components of a semantic analysis which may heavily influence the final output. I am now referring to two well-known cases discussed in the literature: the presence of "conditional" discourse markers like IF, WHETHER which transform a statement into a conditional clause which is usually accompanied by the presence of "unreal" mood like conditional or subjunctive. And then we come to "comparative" constructions which are more frequent in consumer product reviews than in blogs or social networks opinions. As far as comparatives are concerned, it is a fact that real utterances contain a gamut of usage of such a construction which is very hard to come to terms with. We list some of the most relevant cases here below and then make some comments. Each utterance is taken from Bing’s interview databases and has an evaluation at the beginning of the line:

a. *Positive-2 player - i did not want to have high expectations for this apex player because of the price but it is definitely working out much better than what i would expect from an expensive high-end player .

b. *Positive-2 look - without a doubt the finest looking apex dvd player that i 've seen .

c. *Positive-2 dvd player - so sit back , relax and brag to all your friends who paid a mountain of money for a dvd player that can't do half the things this one can , and for a fraction of the price !

d. *Positive-3 camera - recent price drops have made the g3 the best bargain in digital cameras currently available .

e. *Positive-2 feel - you feel like you are holding something of substance , not some cheap plastic toy .

f. *Positive-3 camera - i can't write enough positive things about this great little camera !

g. *Positive-3 camera - this is my first digital camera and i couldn't be happier .

h. *Positive-3 finish - its silver magnesium finish is stunning , and the sharp lines and excellent grip are better than any other camera i've seen .

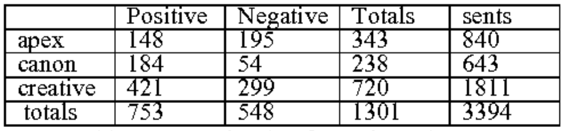
i. *Positive-2 noise another good thing is that this camera seems to introduce much less noise in dark places than others i've seen .

k. *Positive-2 camera this is by far the finest camera in its price and category i have ever used.

As can be noticed, in many cases what is really the guiding principle is the need of comparing the evaluative content of two opposing propositions, rather than simply measuring degree of comparison (superlative rathen than comparative grade). In example a. the first proposition is negated and then the second compared proposition marked by BUT is a really hard complex sentence to compute. In b. one has to compute correctly “without a doubt”. In c. the first proposition has a relative clause referring to a negative fact, where however the governing verb BRAG can be understood both negatively and positively. In d. the phrase “recent price drops” can be a negative fact but has to be understood positively together with the following proposition where “best bargain” appears. Again in e. one needs to compare two propositions one of which has an ellipsed VP. In f. the reviewer uses a rhethorical devise “can’t write enough positive…” which however introduces negation. The same applies to example g.

2.7 The experiment

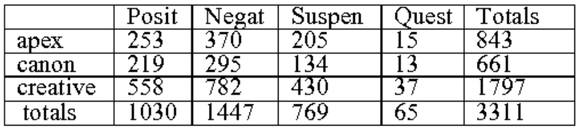
In order to evaluate the system, we used Bing’s (2004) datasets which have been collected and partially annotated in 2004 and are constituted by customer reviews of 5 products downloaded from Amazon.com. In fact we used only three of them – Canon (digital camera), Creative (mp3 player) and Apex (dvd player) - for perusal and for evaluation: the question was that the annotated examples were just a small percentage - 1302 sentences over 3300, so we had to annotate the remaining cases (60% of all utterances) ourselves and make some corrections: the texts were full of typos and had many nonwords, fragments, ungrammatical sentences etc. Overall, we parsed 30,000 tokens and 3300 utterances.



	Positive	Negative	Totals	sents
apex	148	195	343	840
canon	184	54	238	643
creative	421	299	720	1811
totals	753	548	1301	3394

Table 1. Annotation data from Bing's datasets

In Table 1. we report annotation data from the three datasets we used in our experiment, where under SENTS we indicate the number of total utterances present in each dataset. As can be easily gathered, only 38.34% of all utterances have been annotated, which makes the comparison fairly difficult to draw. In particular, if we look at our annotation data below, the overall number of NEGATIVE polarity judgements constitute 58% of all judgements when compared to 42% in Bing's annotation. The final outcome is then totally mistaken: in our case the judgements are more negative and in Bing's they are more positive disregarding each separate product. We computed the number of annotations in Bing's datasets which have been graded [+/- 1], thus indicating that the confidence of the annotator is very low, and this makes up 16.37% of all annotations. In our case, the SUSPENSION annotations constitute 23.22% of all annotations.



	Posit	Negat	Suspen	Quest	Totals
apex	253	370	205	15	843
canon	219	295	134	13	661
creative	558	782	430	37	1797
totals	1030	1447	769	65	3311

Table 2. Evaluation results from Venses

The first interesting data to notice is the slight difference in Recall, where we see that of all the utterances present we only got 97.55%. It is important to highlight the difference in the approach. Our system's output refers real utterances which sometimes do not coincide with each line or record in the input file. The system computes an utterance everytime it finds a sentence delimiting punctuation mark. As a result, in some cases, as in "canon" dataset, we end up with additional utterances to evaluate.

	Fact	Opin	Opin_ Inter	Fact/Opin	Total
apex	398	265	87	100	850
canon	300	226	47	75	648
creative	772	609	195	219	1795
totals	1470	1100	329	394	3293

Table 3. Evaluation results from Venses

Data related to SUBJECTIVITY and FACTIVITY show a balanced subdivision of all data between the two categories.

References

Delmonte R., (2007), *Computational Linguistic Text Processing – Logical Form, Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York.

Delmonte, R.: Text Understanding with GETARUNS for Q/A and Summarization, Proc. ACL 2004 - 2nd Workshop on Text Meaning & Interpretation, Barcelona, Columbia University (2004) 97-104.

Delmonte R., et al. Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach. In: ROMAND 2006 - 11th EACL. Geneva, (2006) 3-10.

Delmonte R., 2008. Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution, Nova Science Publishers, New York.

X. Ding and B. Liu. The Utility of Linguistic Rules in Opinion Mining." *SIGIR-2007* (poster paper).

A. Esuli and F. Sebastiani, EACL-06, 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining, EACL-06, 2006.

C. Fellbaum. *WordNet: an Electronic Lexical Database*, MIT Press, 1998.

M. Ganapathibhotla, B.Liu. Mining Opionions in Comparative Sentences, *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported license,* 2008.

M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD'04*, 2004.

N. Jindal, and B. Liu. Mining Comparative Sentences and Relations. In *AAAI'06*, 2006.

N. Kaji and M. Kitsuregawa. Automatic Construction of Polarity-Tagged Corpus from HTML Documents. *COLING/ACL'06*, 2006.

H. Kanayama and T. Nasukawa. Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis. *EMNLP'06*, 2006.

S. Kim and E. Hovy. Determining the Sentiment of Opinions. *COLING'04*, 2004.

S. Kim and E. Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. *COLING/ACL 2006.*

N. Kobayashi, R. Iida, K. Inui and Y. Matsumoto. Opinion Mining on the Web by Extracting Subject-Attribute-Value Relations. In *Proc. of AAAI-CAAW'06*, 2006.

Pang Bo, L.Lee. 2008. Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2, 1–135.

T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? Finding strong and weak opinion clauses. *AAAI'04*, 2004.

J. Wiebe, and R. Mihalcea. Word Sense and Subjectivity. In *ACL'06*, 2006.
J. Wiebe, and E. Riloff: Creating Subjective and Objective sentence classifiers from unannotated texts. *CICLing*, 2005.
http://www.wjh.harvard.edu/~inquirer/homecat.htm (General Inquirer).
http://sentiwordnet.isti.cnr.it/

Linguistically-based Reranking of Google's Snippets with GreG

Rodolfo Delmonte, Rocco Tripodi

Department of Language Science
Università "Ca Foscari"
30123 – Venezia, Italy

ABSTRACT

We present an experiment evaluating the contribution of a system called GReG for reranking the snippets returned by Google's search engine in the 10 hits presented to the user and captured by the use of Google's API. The evaluation aims at establishing whether or not the introduction of deep linguistic information may improve the accuracy of Google or rather it is the opposite case as maintained by the majority of people working in Information Retrieval and using a Bag Of Words approach. We used 900 questions and answers taken from TREC 8 and 9 competitions and execute three different types of evaluation: one without any linguistic aid; a second one with tagging and syntactic constituency contribution; another run with what we call Partial Logical Form. Even though GReG is still work in progress, it is possible to draw clear cut conclusions: adding linguistic information to the evaluation process of the best snippet that can answer a question improves enormously the performance. In another experiment we used the actual texts associated to the Q/A pairs distributed by one of TREC's participant and got even higher accuracy.

1. Introduction

We present an experiment run using Google API and a fully scaled version of GETARUNS, a system for text understanding (see Delmonte 2007; 2005), together with a modified algorithm for semantic evaluation presented in RTE3 under the acronym of VENSES (Delmonte 2007). The aim of the experiment and of the new system that we called GReG (GETARUNS ReRANKS Google), is that of producing a reranking of the 10 hits presented by Google in the first page of a web search. Reranking is produced solely on the basis of the snippets associated to each link – two per link.

GReG uses a very "shallow" linguistic analysis which nonetheless ends up with a fully instantiated sentence level syntactic constituency representation, where grammatical functions have been marked on a totally bottom-up analysis and the subcategorization information associated to each governing predicate – verb, noun, adjective. More on this process in the sections below.

At the end of the parsing process, GReG produces a translation into a flat minimally recursive Partial Logical Form (hence PLF) where besides governing predicates – which are translated into corresponding lemmata – we use the actual words of the input text for all linguistic relations encoded in the syntactic structure. Eventually all matching processes are carried out coupling semantic similarity measures over the words involved, dependency labels and logical relations.

The evaluation will focus on a subset of the questions used in TREC made up of 900 question/answers pairs made available by NIST and produces the following data:

- how many times the answer is contained in the 10 best candidates retrieved by Google;
- how many times the answer is ranked by Google in the first two links – actually we will be using only snippets (first two half links);
- as a side-effect, we also know how many times the answer is not contained in the 10 best candidates and is not ranked in the first two links;
- how many times GReG finds the answer and reranks it in the first two snippets;
- how much contribution is obtained by the use of syntactic information;
- how much contribution is obtained by means of LF, which works on top of syntactic representation;
- how much contribution is obtained by modeling the possible answer from the question, also introducing some meta operator – we use OR and the *.

The metric we adopt is very similar to the one proposed in Bouma et al. 2005, where they use what they call d-score for dependency relations evaluation, and t-score for syntactic dependency evaluation. The additional information we compute is related to the way we match head words or predicates, which are checked not only for equivalence but also for semantic similarity using the set of semantic relations made available by WordNet; the two words may also belong to the same semantic field as computed by Roget's Thesaurus and other similar lexical resources. Eventually, we compute accuracy measures by means of the usual Recall/Precision formula.

1.1 State of the Art

There is now general consensus on the usefulness of linguistic processing for Q/A open/closed domain tasks. However, the need to keep the processing to a feasible amount of CPU time has led many researchers into the idea that dependency parsing is the only technology able to cope with the task. In fact, dependency word level parsing can become a too poor linguistic representation in many relevant cases. Some of the problems have been overcome by introducing "equivalence paraphrases" which are used to account for syntactic variations (Wang et al. 2007; Bouma et al. 2005).

Some other problems are more related to semantic completeness and factitivity. We are referring here to the problem of recovery of implicit arguments, either by means of binding of syntactic variables or by attaching the appropriate label to underlying object of passive sentences, or again binding the subject of untensed clauses, like gerundives, participials or infinitives. The other problem is related to the need to account for the presence of modality and negation operators which may affect the truth of the answer recovered and thus jeopardize the correctness of the results. These problems are coped with at the level of text entailment evaluation.

In our system, we address different levels of representations – syntactic and (quasi) logical/semantic, and measure their contribution if any in comparison to a baseline keyword or bag of words computation. Together with linguistic representation, we also use semantic similarity evaluation techniques already introduced in RTE challenges which seem particularly adequate to measure the degree of semantic similarity and also semantic consistency or non-contradictoriness of the two linguistic descriptions to compare. This is partially also proposed by others (Wang et al. 2007) when they introduce the use of WordNet to do answer expansion.

Differently from the majority of the systems in this field, we don't use any training, nor do we adopt a specific statistical model. The reason being simply the fact that we want our system to be highly performing in every situation and this may be only guaranteed by a solid and robust linguistic architecture.

2. The Parser

The architecture of the parser is commented in this section. It is a quite common pipeline: all the code runs in Prolog and is made up of manually built symbolic rules.

We defined our parser "mildly bottom-up" because the structure building process cycles on a procedure that collects constituents. This is done in three stages: at first chumks are built around semantic heads – verb, noun, adjective, adverbials. Then prepositions and verb particles are lumped together. In this phase, also adjectives are joined to the nominal head they modify. In a third phase, sentential structure information is added at all levels – main, relative clauses, complement clauses. In presence of conjunctions, different strategies are applied according to whether they are coordinating or subordinating conjunctions.

An important linguistic step is carried out during this pass: subcategorization information is used to tell complements – which will become arguments in the PLF – and adjuncts apart. Some piece of information is also offered by linear order: SUBJect NPs will usually occur before the verb and OBJect NP after. Constituent labels are then substituted by Grammatical Function labels. The recursive procedure has access to calls collecting constituents that identify preverbal Arguments and Adjuncts including the Subject if any: when the finite verb is found the parser is hampered from accessing the same preverbal portion of the algorithm and switches to the second half of it where Object NPs, Clauses and other complements and adjuncts may be parsed. Punctuation marks are also collected during the process and are used to organize the list of arguments and adjuncts into tentative clauses.

The clause builder looks for two elements in the input list: the presence of the verb-complex and punctuation marks, starting from the idea that clauses must contain a finite verb complex: dangling constituents will be adjoined to their left adjacent clause, by the clause interpreter after failure while trying to interpret each clause separately.

The clause-level interpretation procedure interprets clauses on the basis of lexical properties of the governing verb. This is often non available in snippets. So in many cases, sentence fragments are built.

If the parser does not detect any of the previous structures, control is passed to the bottom-up/top-down parser, where the recursive call simulates the subdivision of structural levels in a grammar: all sentential fronted constituents are taken at the CP level and the IP (now TP) level is where the SUBJect NP must be computed or else the SUBJect NP may be in postverbal position with Locative Inversion structures, or again it might be a subjectless coordinate clause. Then again a number of ADJuncts may be present between SUBJect and verb, such as adverbials and parentheticals. When this level is left, the parser is expecting a verb in the input string. This can be a finite verb complex with a number of internal constituents, but the first item must be definitely a verb. After the (complex) verb has been successfully built, the parser looks for complements: the search is restricted by lexical information. If a copulative verb has been taken, the constituent built will be labelled accordingly as XCOMP where X may be one of the lexical heads, P,N,A,Adv.

The clause-level parser simulates the sentence typology where we may have verbal clauses as SUBJect, Inverted postverbal NPs, fronted that-clauses, and also fully inverted OBJect NPs in preverbal position.

2.1 Parsing and Robust Techniques

The grammar is equipped with a lexicon containing a list of fully specified inflected word forms where each entry is followed by its lemma and a list of morphological features, organized in the form of attribute-value pairs. However, morphological analysis for English has also been implemented and used for Out of Vocabulary (hence OOV) words. The system uses a core fully specified lexicon, which contains approximately 10,000 most frequent entries of English. Subcategorization is derived from FrameNet, VerbNet, PropBank and NomBank. These are all consulted at runtime. In addition to that, there are all lexical forms provided by a fully revised version of COMLEX. In order to take into account phrasal and adverbial verbal compound forms, we also use lexical entries made available by UPenn and TAG encoding. Their grammatical verbal syntactic codes have then been adapted to our formalism and is used to generate an approximate subcategorization scheme with an approximate aspectual and semantic class associated to it – some information is derived from LCS from the University of Maryland. Semantic inherent features for OOV words, be they nouns, verbs, adjectives or adverbs, are provided by a fully revised version of WordNet – 270,000 lexical entries - in which we used 75 semantic classes similar to those provided by CoreLex. In addition to that we have a number of gazetteers and proper nouns lists including Arabic names, which amount to an additional 400,000 fully classified lexical entries.

3. Default Semantic Matching Procedure

As said above, the idea is to try to verify whether deeper linguistic processing can contribute to question answering. As will be shown in the following tables, Google's search on the web has high recall in general: almost 90% of the answers are present in the first ten results presented to the user. However, we wanted to assume a much stricter scenario closer in a sense to TREC's tasks. To simulate a TREC task as close as possible we decided that only the first two snippets – not links - can be regarded a positive result for the user. Thus, everything that is contained in any of the following snippets will be computed as a negative result. We take two snippets to be similar to one sentence which is then the basis for the actual answer string.

The decision to regard the first two snippets as distinctive for the experiment is twofold. On the one side we would like to simulate as close as possible a TREC Q/A task, where however rather than presenting precise answers, the system is required to present the sentence/snippet containing it. The other reason is practical or empirical and is to keep the experiment user centered: user's attention should not be forced to spend energy in a tentative search for the right link. Focussing attention to only two snippets and two links will greatly facilitate the user. In this way, GReG could be regarded as an attempt at improving Google's search strategies and tools.

In order to evaluate the contribution of different levels of computation and thus get empirical evidence that a deep linguistically-based approach is worth while trying, we organized the experiment into a set of concentric layers of computation and evaluation as follows:

- at the bottom level of computation we situated what we call the "default semantic matching procedure". This procedure is used by all the remaining higher level of computation and thus it is easy to separate its contribution from the overall evaluation;

- the default evaluation takes input from the first two processes, tokenization & multiword creation plus sentence splitting. Again these procedures are quite standard and straightforward to compute. So we want to assume that the results are easily reproducible as well as the experiment itself;

- the following higher level of computation may be regarded partly system dependent, but most of it is easily reproducible using off-the-shelf algorithms made available for English by research centers all over the world. It regards tagging and context-free PennTreebank-like phrase-structure syntactic representation as well as dependency parsing. Here we consider not only words, but word-tag pairs and word-as-head of constituent N pairs. We also take into account their grammatical function label;

- the highest level is constituted by what we call Partial Logical Form, which builds a structure containing a Predicate and a set of Arguments and Adjuncts each headed by a different functor. In turn each such structure can contain Modifiers. Each PLF can contain other PLFs recursively embedded with the same structure. More on this below. This can also be reproduced by algorithms available off-the-shelf at the DELPH-IN website.

3.1 A walkthrough example

We now present three examples taken from TREC8 question/answer set, no. 3, 193, 195, corresponding respectively to ours 1,2,3. For each question we add the answer and then we show the output of tagging in PennTreebank format, then follows our enriched tagset and then the syntactic constituency structure produced by the parser and the grammatical labels. Eventually, we show the Partial Logical Form where the question word has been omitted. The question word will be transformed into its corresponding semantic type. In some cases, it can be reinserted in the analysis when the matching takes place and may appear in the other level of representation we present which is constituted by the Query in answer form passed to Google. Question words are always computed as argument or adjunct of the main predicate, so GReG will add a further match with the input snippets constituted by the semantic types of the wh- words. One such type is visible in question no.3 when the concept "AUTHOR" is automatically added by GReG in front of the verb and after the star. More on answer typing below.

(1) What does Peugeot company manufacture? – Cars

(2) Who was the 16th President of the United States? – Lincoln

(3) Who wrote "Dubliners"? – James Joyce

Here below are the analyses where we highlight the various levels of linguistic representation relevant for our experiment only – except for the default word level:

(1) Tagging and Syntactic Constituency

what-wp, does-md, the-dt, Peugeot-nnp, company-nn, manufacture-vin, ? – pun

[what-int, does-vsup, the-art, Peugeot-n, company-n, manufacture-vin, ? - puntint]

cp-[cp-[what], f-[subj-[the, company, mod-[Peugeot]], ibar-[does, manufacture]], fint-[?]]

Partial Logical Form
pred(manufacture)arg([company, mod([Peugeot])]) adj([[], mod([[]])])

Query launched to Google API
Peugeot company manufacture *

(2) Tagging and Syntactic Constituency
who-wp, was-vbd, the-dt, 16th-cd, President-nnp, of-in, the-dt, United_States-nnp, ? – pun

[who-int, was-vc, the-art, 16th-num, President-n, of-p, the-art, United_States-n, ? - puntint]

fint-[cp-[who], ibar-[was], sn-[the, 16th, President, mod-[of, the, United_States]], fint-[?]]

Partial Logical Form
[pred(be) arg([President, mod([united, States, 16th])]) adj([])]

Query launched to Google API
United States 16th President was *

(3) Tagging and Syntactic Constituency
who-wp, wrote-vbd_vbn, "-pun, Dubliners-nns, "-pun, ? - pun

[who-int, wrote-vt, "-par, Dubliners-n, "-par, ? - puntint]

cp-[cp-[who], ibar-[wrote], fp-["], sn-[Dubliners], fp-["], fint-[?]]

Partial Logical Form
pred(write) arg([Dubliners, mod([])]) adj([])

Query launched to Google API
* author wrote Dubliners

3.2 Default Semantic Matching Procedure

This is what constitutes the closest process to the BOWs approach we can conceive of. We compare every word contained in the Question with every word contained in each snippet and we only compare content words. Stopwords are deleted.

We match both simple words and multiwords. Multiwords are created on the basis of lexical information already available for the majority of the cases. The system however is allowed to guess the presence of a multiword from the information attached to the adjacent words and again made available in our dictionaries. If the system recognizes the current word as a word starting with uppercase letter and corresponding to one of the first names listed in one of our dictionary, it will try to concatenate this word to the following and try at first a match. If the match fails the concatenated word is accepted as a legitimate continuation – i.e. the name – only in case it starts by uppercase letter. Similar checking procedures have been set up for other NEs like universities, research centres, business related institutions etc. In sum, the system tries to individuate all NEs on the basis of the information stored and some heuristic inferential mechanism.

According to the type of NE we will licence a match of a simple word with a multiword in different ways: person names need to match at least the final part of the multiword, or the name institutions, locations etc. need to match as a whole.

3.3 Tags and Syntactic Heads

The second level of evaluation takes as input the information made available by the tagger and the parser. We decided to use the same approach reported in the challenges called RTE where the systems participating could present more than one run and use different techniques of evaluation. The final task was – and is – that of evaluating the semantic similarity between the question and the input snippets made available by Google. However, there is a marked difference to be taken into account and is the fact that in RTE questions where turned into a fully semantically complete assertion; on the contrary, in our case we are left with a question word – applies to wh- questions - to be transformed into the most likely linguistic description that can be associated with the rest of the utterance. As most systems participating in TREC challenge have done, the question has to be rephrased in order to predict the possible structure and words contained in the answer, on the basis of the question word and overall input utterance. Some of the questions contained in the TREC list do not actually constitute wh- questions (factoid or list), but are rather imperatives or iussive utterance, which tell the system – and Google – to "describe" or to "name" some linguistic item specified in the following portion of the utterance.

As others have previously done, we classify all wh- words into semantic types and provide substitute words to be place in the appropriate sentence position in order to simulate as close as possible the answer. In other cases the semantic type will be used to trigger the appropriate general concept associated to the corresponding word matched in the snippet. In particular, whenever a number is required, be it a date, or other, the type QUANTITY is used to trigger the appropriate type. We distinguish between: DISTANCE, REPEAT, DURATION, DATE, POPULATION and a generic QUANTITY. In the latter case, however, a specific procedure checks for special cases of quantity definition, which may be SHARP, LESS, MORE, ABOUT, INCLUDES. Each subtype will compute the similarity accordingly.

However, this is only done in one of the modalities in which the experiment has been run. In the other modality, Google receives the actual words contained in the question.

As to experiment itself, and in particular to the matching procedure we set up, the wh- word is never used to match with the snippets. Rather we use the actual wh- words to evaluated negatively snippets containing them. In this way, we prevent similar and identical questions contained in a snippet and pointed by a link to receive a high score. We noticed that Google is unable to detect such mismatches.

We decided to use tag-word pairs in order to capture part of the contextual meaning associated to a given word. Also in the case of pairs word-as-head-of-constituent/ constituent label we wanted to capture part of the contextual import of a word in a structural representation and thus its syntactic and semantic relevance in the structure. As will be clear in the following section, this is different from what is being represented in a Logical Form for how partial it may be.

3.3 Partial Logical Form and Relations

The previous match intended to compare words as part of a structure of dependencies where heads played a more relevant role than non-heads, and thus were privileged. In the higher level match what we wanted to check was the possible relations intervening between words: in this case, matching regarded two words at a time in a hierarchy. The first and most relevant word was the PREDicate governing a given piece of PLF. The PRED can be the actual predicate governing at sentence level, with arguments and adjuncts, or it can be just the predicate of any of the Arguments/Adjuncts which in turn governs their modifiers.

Matching is at first applied to two predicates and if it succeeds, it is extended to the contents of the Argument or the Adjunct. In other words, if it is relations that this evaluation should measure, any such relations has to involve at least two linguistic elements of the PLF representation under analysis.

Another important matching procedure applied to the snippet is constituted by a check of the verbal complex. We regard the verbal compound as the carrier of semantically important information to be validated at propositional level. However, seen the subdivision of tasks, we assume that we can be satisfied by applying a partial match. This verbal complex match is meant to ascertain whether the question and the answer contain positive polarity items; and in case they contain negative polarity items then they should be both containing one such item –not to convey contradictory information. It is also important to check whether the two verbal complexes are factitive or not: this is checked by detecting the presence of opaque or modality operators in the verb complex and at propositional adjunct level. This second possibility is matched carefully.

4. Evaluation

Here below we show the output of GReG in relation to one of the three questions presented above, question n.2

```
google7
Evaluation Score from Words and Tags : 31
Evaluation Score from Syntactic Constituent-Heads : 62
Evaluation Score from Partial Logical Form : 62
google8
Evaluation Score from Words and Tags : 35
Evaluation Score from Syntactic Constituent-Heads: 70
Evaluation Score from Partial Logical Form :  0
google9
Evaluation Score from Words and Tags : 33
Evaluation Score from Syntactic Constituent-Heads : 66
Evaluation Score from Partial Logical Form : 66

Snippet No.  google9
16th President of the United States ( March 4 , 1861 to April 15 , 1865 ).
Nicknames : “ Honest Abe “ “ Illinois Rail - Splitter “. Born : February 12
, 1809 , . . .

Snippet No.  google7
Abraham Lincoln , 16th President of the United States of America , 1864 ,
Published 1901 Giclee Print by Francis Bicknell Carpenter - at AllPosters
. com .
```

The right answer is : Lincoln

Google's best snippets containing the right answer are:

google8
Who was the 16th president of the united states ? pissaholic . . . . Abraham Lincoln was the Sixteenth President of the United States between 1861 - 1865 . . .

google7
Abraham Lincoln , 16th President of the United States of America , 1864 , Published 1901 Giclee Print by Francis Bicknell Carpenter - at AllPosters . com .

Google's best answer partially coincides with GReG.

****

Passing Questions to Google filtered by GReG’s analysis produced a positive result in that 755 questions contained the answer in the 10 best links. On the contrary, passing Questions to Google as is, produces as a result that only in 694 questions contain the answer in the 10 best links. In other words, GReG’s analysis of the question triggers best results from Google, in fact improving the ability of Google to search for the answer and select it in the best 10 links.

In fact, Google exploits the linear order of words contained in the question. So in case there is some mismatch the answer is not readily found or perhaps is available further down in the list of links.

	With GReG’s preanalysis	Without GReG’s anal.
Google’s 10 Best links contain the answer	755 83.01%	694 77.12%
Google’s 10 Best links do not contain the answer	145 16.9%	206 22.8%
Google Rank answer in first 2 snippets	216 28.61%	168 24.21%
Google Rank answer not in first 2 snippets	684 76.00%	732 82.34%

Table 1: Google outputs with and without the intervention of GReG’s question analysis

GReG reranks the answer in	Only word match	Tagging and Syntactic	Partial Logical Form

first 2 snippets		heads	
With GReG's analysis	375 58.41%	514 68.08%	543 71.92%
Without GReG's analysis	406 55.09%	493 66.89%	495 67.16%

Table 2: GReG's outputs at different levels of linguistic complexity

4.2 Discussion

The conclusions we may safely draw is the clear improvements in performance of the system when deep linguistic information is introduced in the evaluation process. In particular, when comparing the contribution of PLF to the reranking process we see that there is a clear improvement: in the case of reranking without GReG's question analysis there is a slight but clear improvement in the final accuracy. Also, when GReG is used to preanalyse the question to pass to Google the contribution of PLF is always apparent. The overall data speak in favour of both preanalysing the question and using more linguistic processing.

If we consider Google's behaviour to the two inputs, the one with actual questions and the one with prospective answers we see that the best results are again obtained when the preanalysis is used; also the number of appropriate candidates – the recall - containing the answer increases remarkably when using GReG preprocessing (83% vs. 77%).

4.3 GReG and Question-Answering from Text

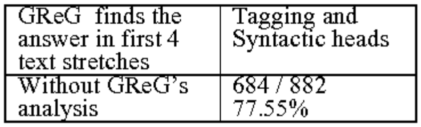
In order to verify the ability of our system to extract answers from real text we organized an experiment which used the same 900 question run this time against the texts made available by TREC participants. These texts have two indices at the beginning of each record line indicating respectively the question number which they should be able to answer, and the second an abbreviation containing the initial letters of the newspaper name and the date. In fact each record has been extracted by means of automatic splitting algorithms which have really messed up the whole text. In addition, the text itself has been manipulated to produce tokens which however do not in the least correspond to actual words of current orthographic forms in real newspapers. So it took us quite a lot of work to normalize the texts (5Mb.) to make them as close as possible to actual orthography.

Eventually, when we launched our system it was clear that the higher linguistic component could not possibly be used. The reason is quite simple: texts are intermingled with lists of items, names and also with tables. Since there is no principled way to tell these apart from actual texts with sentential structure, we decided to use only tagging and chunking.

We also had to change the experimental setup we used with Google snippets: in this case, since we had to manipulate quite complex structures and the choice was much more noisy, we raised our candidate set from two to four best candidates. In particular we did the following changes:

- we choose all the text stretches – usually corresponding to sentences - containing the answer/s and ranked them according to their semantic similarity;
- then, we compared and evaluated these best choices with the best candidates produced by our analyses;
- we evaluated to success every time one of our four best candidates was contained in the set of best choices containing the answer;
- otherwise we evaluated to failure.

In total, we ran 882 questions because some answers did not have the corresponding texts. Results obtained after a first and only run – which took 1 day to complete on an HP workstation with 5GB of RAM, 4 Dual Core Intel processors, under Linux Ubuntu – were quite high in comparison with the previous ones, and are reported here below:



GReG finds the answer in first 4 text stretches	Tagging and Syntactic heads
Without GReG's analysis	684 / 882 77.55%

Table 3: GReG's results with TREC8/9 texts

With respect to the favourable results, we need to consider that using texts provides a comparatively higher quantity of linguistic material to evaluate and so it favours better results.

5. Conclusions and future work

We intend to improve both the question translation into the appropriate format for Google, and the rules underlying the transduction of the Syntactic Structures into a Partial Logical Form. Then we will run the experiments again. Considering the limitations imposed by Google on the total number of questions to submit to the search engine per day, we are unable to increase the number of questions to be used in a single run.

We also intend to run GReG version for text Q/A this time with question rephrasing. We would also like to attempt using PLF with all the text stretches, after excluding manually all tables and lists. We are aware of the fact that this would constitute a somewhat contrived and unnatural way of coping of unrestricted text processing. At the same time we need to check whether the improvements we obtained with snippets are attested by the analysis of complete texts.

Overall, we believe to have shown the validity of our approach and the usefulness of deep linguistically-based evaluation methods when compared with shallower approaches. Structural and relational information constitutes a very powerful addition to simple tagging or just word level semantic similarity measures.

References

Bouma G., Mur J., van Noord G., 2005. Reasoning over dependency relations", Proceedings of KRAQ.

Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, Tat-Seng Chua, 2005. Question Answering Passage Retrieval Using Dependency Relations, SIGIR'05, ACM, Salvador, pp.400-406.

Delmonte R., 2007. Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering, Nova Science Publishers, New York.

Delmonte R., 2005. Deep & Shallow Linguistically Based Parsing, in A.M.Di Sciullo(ed), UG and External Systems, John Benjamins, Amsterdam/Philadelphia, pp.335-374.

Delmonte R., A. Bristot, M.A.Piccolino Boniforti, S.Tonelli 2007. Entailment and Anaphora Resolution in RTE3, in Proc. ACL Workshop on Text Entailment and Paraphrasing, Prague, ACL Madison, USA, pp. 48-53.

Litkowski, K. C. 2001. Syntactic Clues and Lexical Resources in Question-Answering. In E. M. Voorhees & D. K. Harman (eds.), The Ninth Text Retrieval Conference (TREC-9). NIST Special Publication 500-249. Gaithersburg, MD., 157-166.

Mengqiu Wang and Noah A. Smith and Teruko Mitamura, 2007. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA, in Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp. 22-32.

Sentiment Analysis of French Movie Reviews

Hatem Ghorbel and David Jacot

Abstract In sentiment analysis of reviews we focus on classifying the polarity (positive, negative) of conveyed opinions from the perspective of textual evidence. Most of the work in the field has been intensively applied on the English language and only few experiments have explored other languages. In this paper, we present a supervised classification of French movie reviews where sentiment analysis is based on some shallow linguistic features such as POS tagging and word semantic orientation extracted from the lexical resource SentiWordNet. Since SentiWordNet is an English resource, we apply a word-translation from French to English before polarity extraction. We show moreover in this article the problems derived by such a translation and their consequences on the word semantic orientation.

1.1 Introduction

Sentiment analysis is an emerging discipline whose goal is to analyze textual content from the perspective of the opinions and viewpoints they hold. A large number of studies have focused on the task of defining the polarity of a document which is by far considered as a classification problem: decide to which class a document is attributed; class of positive or negative polarity.

Most of the work in the field has been intensively applied on the English language. For this purpose, English resources and corpora (such as MPQL [WWH05], Movie Review Data [PLV02], Product Review [YNBN03], Book Review [GA05], SentiWordNet [ES06], WordNet-Affect [SV04, VSS04], the Whissell's Dictionary

Hatem Ghorbel
University of Applied Sciences Western Switzerland, Haute Ecole Arc Ingénierie, St-Imier, Suisse
e-mail: hatem.ghorbel@he-arc.ch

David Jacot
University of Applied Sciences Western Switzerland, Haute Ecole Arc Ingénierie, St-Imier, Suisse
e-mail: david.jacot@master.hes-so.ch

of Affect Language [Whi89]) have been constructed to aid in the process of automatic supervised and unsupervised polarity classification of textual data. Nevertheless, still very few experiments are applied on other languages.

In this context, we address in this paper the issue of polarity classification but applied on French movie reviews. We used a supervised learning approach where we trained the classifier on annotated data of French movie reviews extracted from the web. As classification features, beyond the word unigrams feature taken as the baseline in our experiments, we extracted further linguistic features including lemmatized unigrams, POS tags and semantic orientation of selected POS tags. The latter feature is extracted from the English lexical resource SentiWordNet after applying a word-translation from the French to English.

The main goal of our experiments is firstly to confirm that the incorporation of linguistic features into the polarity classification task could significantly improve the results. Secondly, to address the problem of loss of precision in defining the semantic orientation of word unigrams from English lexical resources, mainly due to the intermediate process of word-translation from French to English correlated with further issues such as sense disambiguation.

In the rest of the paper, we first shortly describe the previous work in the field of sentiment analysis and polarity classification. Then we describe the set of extracted features used in polarity classification of French movie reviews. Finally we provide and discuss the obtained experiment results and end up by drawing some conclusions and ideas for future work.

1.2 Previous Work

So far, researchers have been used the same classification methodologies and techniques as topic-based categorization [PLV02] with special emphasis on linguistic features in order to increase the performance. As linguistic features, [Gam04, MTO05, NDA06] present syntactically motivated features, most of them based on dependency path information and modeled as high n-grams. Further linguistic features such as part of speech, negation, verbs modality, and semantic information (from WordNet for instance) are recently explored [MPI07, WK09, TNKS09, ABM09].

Moreover, statistical approaches have been coupled with semantic approaches in order to achieve better results [KH04, PL04, WWH05, Osh08]. Semantic approaches aim at classifying sentiment polarity conveyed by textual data using commonsense, sentiment resources, as well as linguistic information. For instance, [HL04, ES05, NSS07, Den08] classify polarity using emotion words and semantic relations from WordNet, WordNet Gloss, WordNet-Affect and SentiWordNet respectively.

An important theoretical issue in the semantic approach is still how to define the semantic orientation of a word in its context. Some studies showed that restricting features to those adjectives would improve performance. [HM97] have focused on defining the polarity of adjectives using indirect information collected from a large corpus. However, more researches showed that most of the adjectives and adverbs, a small group of nouns and verbs possess semantic orientation [TTC09, AB06, ES05, GA05, MTO05, TL03].

Only very few work [Den08, ACS08] have explored sentiment analysis in a multilingual framework such as Arabic, Chinese, English, German and Japanese. Their methodology is based on standard translation from target language to English in order to reuse existing English corpora and resources for polarity classification.

1.3 Feature Design

Similarly to previous sentiment analysis studies, we have defined three categories of features. These include lexical, morpho-syntactic and semantic (word polarity) features. Lexical and morpho-syntactic features have been formulated at the word level, whereas semantic features have been formulated at the review level.

1.3.1 Lexical features

This is the baseline of our experiments and is mainly composed of word unigrams. The global assumption in this choice is that we tend to find certain words in positive reviews and others in negative ones. Each unigram feature formulates a binary value indicating the presence or the absence of the corresponding word at the review level. In order to improve the relevance of unigram features, we propose below further variants.

Stop words The French language contains a lot of stop words like *"de"*, *"du"*, *"à"*, *"le"* and *"la"*. Generally, these words don't hold polarity information so they aren't relevant for the classifcation. A stop list for removing those words may improve the results.

Lemmatization Grouping all inflected forms of a word in a single term may be usefull in sentiment analysis. For example, consider the words *"aimé"*, *"aimait"*, *"aimer"*, *"aiment"* and *"aime"*, all these words share the same polarity but will be considered as five seperate features during the classification. When applying lemmatization, we would obtain a unique feature. Features reduction would improve the tuning of the training process.

Moreover the lemmatization is quite important for our experiments because WordNet and SentiWordNet use lemmatized words in their dictionary.

1.3.2 Morpho-syntactic features

Definitely not all the words are relevant to the sentiment classification. Some studies showed that restricting features to adjectives would improve performance [HM97], for instance. Part-of-speech (POS) tags are used to add information to word unigrams features in order to disambiguate words that share the same spelling but not the same polarity. For example, it would distinguish the different usages of the word *"négatif"* that can either be a neutral noun (*"un négatif"*) or a negative adjective (*"un commentaire négatif"*). Moreover POS is also important to aid word sense disambiguation before polarity extraction in SentiWordNet.

1.3.3 Semantic features

As it is shown is previous work [HL04, ES05, NSS07, Den08], the incorporation of corpus and dictionary based resources such as WordNetAffect, SentiWordNet and Whissell's Dictionary of Affect Language contributes in improving the sentiment classification. Based on such results, we use the lexical resource SentiWordNet[1] to extract word polarity and calculate the overall polarity score of the review for each POS tag. SentiWordNet is a corpus-based lexical ressource constructed from the perspective of WordNet. It focuses on describing sentiment attributes of lexical entries describe by their POS tag and assigns to each synset of WordNet three sentiment scores: positivity, negativity and objectivity.

Since SentiWordNet describes English lexical resources, we go through a word-translation from French to English before polarity extraction. Words are lemmatized before being passed through the bilingual dictionary. We use POS information as well as the most frequently[2] used sense selection to disambiguate senses and predict the right synset. We only considered the positivity and the negativity features for the four POS tags noun, adjective, verb and adverb for this task.

More specifically, we added for each review and for each POS tag two features holding the scores of negativity and positivity as extracted from SentiwordNet. These two scores are calculated as the sum of polarities over all the words of the review respecting POS categorization.

1.4 Experiments

Since we didn't find any available sets of annotated data (already classified as negative or positive) of French movie reviews, we collected our own data from the web[3].

[1] SentiWordNet 1.0.1

[2] This choice is based on the assumption that reviewers spontaneously use an everyday language.

[3] We extracted spectators' reviews from `http://www.allocine.com`

Table 1.1 Performance of Different Feature Sets.

	Features	# of features	Results [%]		
			Pos.	Neg.	Acc.
(1)	Unigrams	14635	92.00	91.00	91.50
(2)	Unigrams + stop list	14270	92.00	91.50	91.75
(3)	Unigrams + lemmatisation	10624	92.00	93.00	92.50
(4)	Unigrams + lemma. + POS	12229	93.00	92.50	92.75
(5)	Unigrams + lemma. + POS (N, V, ADJ, ADV)	10350	90.00	92.50	91.25
(6)	Unigrams + lemma. + POS (ADJ)	2109	79.50	92.00	85.75
(7)	Unigrams + lemma. + polarity	10632	93.00	93.50	93.25
(8)	Unigrams + lemma. + POS + polarity	12237	92.00	93.50	92.75

We extracted a corpus of 2000 French movie reviews, 1000 positive and 1000 negative, from 10 movies, 1600 were used for training and 400 for testing. We included reviews having a size between 500 and 1000 characters.

Prior classification of the corpus is elaborated according to user scoring: positive reviews are marked between 2.5 and 4 whereas negative reviews are marked between 0 and 1.5[4]. This prior classification is based on the assumption that the scoring is correlated to the sentiment of the review.

For our experiments, the data was preprocessed with the TreeTagger[5], a French POS tagger and lemmatization tool.

We used Support Vector Machine (SVM) classification method to train and classify French movie reviews. We used SVM^{Light} [Joa98] classification tool with its standard configuration (linear kernel) to implement a series of experiments where each time we define a set of combined features and evaluate the accuracy of the approach.

1.4.1 Results and Discussion

The results of the following experiments are summarized in Table 1.1 above. For each experiment labeled from (1) to (8), we present the number of used features and the accuracy mesured on the test corpus.

[4] Scores are bounded between 0 (for very bad) and 4 (excellent) with a step of 0.5. Reviews scored with 2 are not considered in the construction of our corpus since it is hard to manually classify them as positive or negative opinions.

[5] TreeTagger was developed by Helmut Schmid at the Institute for Computational Linguistics of the University of Stuttgart. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

1.4.1.1 Lexical features

Similarly to [PLV02] we encoded all words features as binary features indicating the presence or the absence of a word in a review. As a first step, we included the entire set of words without applying any specific filtration method.

The accuracy (1) using the entire set of words is 91.50%. Comparing this results to [PLV02] reporting an accuracy of 82.90% on English movie reviews classification using similar features, we find that our results are approximately 10% higher. We believe that this gap is due to the nature of our corpus and size of our reviews (the collected French reviews are shorter).

When using a stop list (2) we increase the performance by approximately 0.25% up to 91.75%. Although the size of the feature set is reduce by approximately 2.40%, results are not significantly improved. Indeed, stop words seem to be already ignored by the classifier. Therefore a stop list is not useful.

However using the lemmatization (3), we increase the accuracy by 1.00% up to 92.50% and reduce the feature set by approximately 27.00%. These results are quite encouraging for the rest of our experiments because we need to work with lemmatized unigrams to query SentiWordNet.

In order to understand the misclassification of some reviews, we looked deep in their content and noted the following problems.

Neutral reviews Reviews manually interpreted as neutral such as *"le film est visuellement réussi mais le scénario est d'une banalité affligente"* are randomly classified according to the dominant sentiment of contained words.

Ironic expressions We noticed that ironic expressions such as *"trop fort les gars"* that has a negative polarity although it is composed of positive words.

Negation Some reviews that use negation such as *"il n'y a plus rien d'extraordinaire"* are misclassified. However if we look at the corpus, there are many reviews containing negation and are well classified. For example among about 640 reviews which contain the regex *"ne [a-z]* pas"* there are only 7 misclassified ones. Generally this is a good result.

Prior classification Only one annotation error was found in the test set. Therefore we can consider that our approach for the prior classification works well.

Misspellings Misspellings are not standard unigrams and hence could not regularly be present in the training data. Reviews containing a large number of misspellings would have their features significantly reduced and so provide very poor information for the classification. We noted that isolated and common misspellings don't affect much the classification but reviews which contain relatively many misspellings tend to be misclassified.

Sometimes misspellings could be volontary to express a kind of stress and intonation such as *"énnnorme"*. The problem with such kind of words is that they are irregular in the corpus. For example, *"énnnorme"* is highly positive but it is not present in the feature set so it is not useful. Quite misspelled reviews tend to be misclassified.

Out of scope span Some reviews contain subjective sentences that describe other satellite subjects that do not concern the reviewer opinion about the movie. For example we could find a description of a particular scene that does not necessarily reflect the global opinion about the movie such as *"Monsieur X est très gentil dans le film."*. Such out of scope sentences may affect the classification.

1.4.1.2 Morpho-syntactic features

In further experiments, we appended POS tags to every lemmatized unigram so as to disambiguate same unigrams having different senses. However, the effect of this information seems to be irrelevant, as depicted on line (4) of Table 1.1, the accuracy is only increased by approximately 0.25% up to 92.75%.

When filtering unigrams to retain only nouns, adjectives, adverbes and verbs features, we intuitively expect a better classification since we might assume that polarity of a review is substantially hold by such POS categories. Results, shown on line (5) are disappointing since performance is decreased by 1.25% down to 91.25% comparing to the lemmatized unigrams (3). Furthermore, when restricting unigrams features to only adjectives (6), the performance is getting worse; accuracy is decreased by 6.75% down to 85.75% comparing to (3) and the feature set is reduced by approximately 80%. In order to understand such inconsistency, we look deeper at the accuracy of positive and negative reviews separately. On a one hand, we notice that negative reviews are better classified than positive ones. On the other hand, we found, in additional experiments, that negative reviews contain relatively an important number of positive adjectives (generally in the negative form). Since we didn't take into account the negation, these positive adjectives are assumed to negative features in the training model, which induces a further difficulty when classifying positive reviews containing these positive adjectives. This last experiment is in contradiction with the results of [HM97] but confirms the results of [PLV02].

1.4.1.3 Semantic features

A part from the lexical and the POS features, we extend in our experiments the features set to words polarity extracted from SentiWordNet and formulated as a score representing the overall negativity and positivity of words in the reviews. As shown on the table 1.1, results are improved by only 1.75% up to 93.25% compared to lemmatized unigrams experiment (3). The main reason of such an expectedly and barely perceptible improvement is the failure of extracting polarity information of words from SentiWordNet: among 2000 adjectives, we got the polarity information of only 800 entries in SentiWordNet (40% of success). This extraction problem is mainly due to the following problems.

Translation errors We translate words from French to English to be able to work with SentiWordNet. However, the quality of translation significantly affects the results of semantic polarity extraction and this is due to the following reasons:

- The bilingual translator doesn't preserve the POS of words. For example, the *noun "méchant"* is translated into *"wicked"* which is implicitly an *adjective* and not a *noun*. Since the translator does not reveal information about the POS change after translation, *wicked* is assumed to be a *noun*. However, the *noun "wicked"* doesn't exist in SentiWordNet.
- Moreover, even if the translation is correct, it happens that the two parallel words do not share the same semantic orientation across both languages due to a difference in common usage, for instance the French *positive* adjective *"fééeriques"* is translated into the *negative* English adjective *"magical"*; the French *positive* adjective *"magique"* is translated into the *negative* adjective *"magic"* as found in SentiWordNet.

Lemmatization and POS tagging errors Misspellings are not standard unigrams and hence could not be found in SentiWordNet. Reviews containing a large number of misspellings would have their overall polarity uncorrect. In addition, misspellings and other lexical errors (for example punctuation, use of parenthesis *"permanente(c'est"* and composed words *"a-tu-vu"*) could significantly affect the results of lemmatization and POS tagging tasks elaborated by TreeTagger. In fact, TreeTagger is not implemented to cope with everyday French language as found in spontaneous movie reviews.

Negation As previously mentioned, the negation was not processed. In principle, the polarity of negated words should be inverted: a negative review which contains many positive words in the negative form should show an overall negative polarity and vice-versa.

Adverbs high negativity Some adverbs like *ne*, *pas*, *rien* et *plus* have a very high negativity (about 0.75 over 1). Reviews which abundantly contains these adverbs tend to be classified as negative. We can distinguish two entailments:

- When these words are used in negative reviews, they accentuate the negativity of the document. In fact, this may be useful for negative reviews which relatively contain a large number of positive words in the negation form. Such high negativity may compensate the absence of negation processing and contributes to the improvement of the classification of negative reviews as we have seen in the current experiment.
- However, when these adverbs are used in positive reviews, they tend to inverse the global polarity, mostly for reviews which are weakly positive. In this experiment, some positive reviews are slightly affected by such a problem.

To sum up, we believe that it would be more relevant to eliminate the polarity processing of such adverbs and instead process the negation form.

1.5 Conclusions

In this paper, an unsupervised approach to sentiment analysis of French movie reviews in a bilingual framework was described. It has been shown that the combination of lexical, morpho-syntactic and semantic features achieves relatively good performance in classifying French movie reviews according to their sentiment polarity (positive, negative). Several problems having an effect upon the results of the classification were highlighted and potential solutions were discussed.

In order to extract the semantic orientation of words from SentiWordNet, we went through a standard word-translation process. Although translation does not necessary preserve the semantic orientation of words due to the variation of language common usage especially when it comes to spontaneous reviews on the web, and in spite of all its side effects, it has been argued that dictionary-based approach could contribute to achieve better results. Even if our first experiments showed little significance, further improvements have been proposed accordingly.

In future evaluations, the method will be analyzed within a larger training and test sets. Further linguistic analysis will be elaborated such as misspelling correction, negation, WSD and elimination of out of scope text spans from reviews, in addition to the improvement of the translation task.

References

[AB06] A. Andreevskaia and S. Bergler. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet. *In Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

[ABM09] A. Agarwal, F. Biadsy, and K. McKeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. *In proceedings of the European Chapter of Association for computational Linguistics (EACL-09)*, 2009.

[ACS08] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages:feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26:No. 3, Article 12, June 2008.

[Den08] K. Denecke. Using sentiwordnet for multilingual sentiment analysis. *In proceedings of the IEEE International Conference on Data Engineering (ICDE2008)*, pages 507–512, 2008.

[ES05] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. *In Proceedings of CIKM '05*, pages 617–624, 2005.

[ES06] A. Esuli and F. Sebastiani. Sentiwordnet: a publicly available lexical resource for opinion mining. *In Proceedings of the 5th Conference on Language Resources and Evaluation LREC*, 6, 2006.

[GA05] M. Gamon and A. Aue. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. *In Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, July 2005.

[Gam04] M. Gamon. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. *In Proceedings of the 20th International Conference on Computational Linguistics*, page 611–617, August 2004.

[HL04] M. Hu and B. Liu. Mining and summarizing customer reviews. *In Proceedings of Knowledge Discovery and Data Mining (KDD'04)*, 2004.

[HM97] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. *In Proceedings of the 8th conference on European Chapter of the Association for Computational Linguistics*, page 174–181, 1997.

[Joa98] T. Joachims. Making large-scale svm learning practical. *ACM Transactions on Information Systems (TOIS)*, 1998.

[KH04] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. *In Proceedings of the 20th international conference on computational linguistics (COLING 2004)*, page 1367–1373, August 2004.

[MPI07] S. M. Al Masum, H. Prendinger, and M. Ishizuka. Sensenet: A linguistic tool to visualize numerical-valence based sentiment of textual data. *In Proceedings of the International Conference on Natural Language Processing (ICON)*, page 147–152, 2007.

[MTO05] S. Matsumoto, H. Takamura, and M. Okumura. Sentiment classification using word sub-sequences and dependency-trees. *Lecture notes in computer science*, 3518:301–311, 2005.

[NDA06] V. Ng, S. Dasgupta, and S.M.N Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, page 611–618, July 2006.

[NSS07] V. Nastase, M. Sokolova, and J.S. Shirabad. Do happy words sound happy? a study of the relation between form and meaning for english words expressing emotions. *In Proceedings of Recent Advances in Natural Language Processing (RANLP'2007)*, pages 406–410, 2007.

[Osh08] A. Osherenko. Towards semantic affect sensing in sentences. *In Proceedings of Communication, Interaction and Social Intelligence (AISB-2008)*, 2008.

[PL04] B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL)*, page 271–278, 2004.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, July 2002.

[SV04] C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. *In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086, May 2004.

[TL03] P.D. Turney and M.L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, pages 15–346, 2003.

[TNKS09] T.T Thet, J-C. Na, C. Khoo, and S. Shakthikumar. entiment analysis of movie reviews on discussion boards using a linguistic approach. *In Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion measurement*, 2009.

[TTC09] H. Tang, S. Tan, and X. Cheng. A survey on sentiment detection of reviews. *Expert Systems with Applications: An International Journal*, 36(7):10760–10773, September 2009.

[VSS04] A. Valitutti, C. Strapparava, and O. Stock. Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83, 2004.

[Whi89] C.M. Whissell. The dictionary of affect in language. *lutchik, R., Kellerman, H. (eds.) Emotion: Theory, Research, and Experience*, page 113–131, 1989.

[WK09] M. Wiegand and D. Klakow. The role of knowledge-based features in polarity classification at sentence level. *In Proceedings of the Florida Artificial Intelligence Research Society Conference (FLAIRS Conference 2009)*, 2009.

[WWH05] T. Wilson, J. Wiebe, and P. Homann. Recognizing contextual polarity in phrase-level sentiment analysis. *In Proceeding of the conference on empirical methods in natural language processing (EMNLP 2005)*, page 347–354, October 2005.

[YNBN03] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *In The Third IEEE International Conference on Data Mining*, 2003.

Motivating Serendipitous Encounters in Museum Recommendations

Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Piero Molino

Abstract Recommender Systems try to assist users to access complex information spaces regarding their long term needs and preferences. Various recommendation techniques have been investigated and each one has its own strengths and weaknesses. Especially, content-based techniques suffer of overspecialization problem. We propose to inject diversity in the recommendation task by exploiting the content-based user profile to spot potential surprising suggestions. In addition, the actual selection of serendipitous items is motivated by an applicative scenario. Thus, the scenario concerns with personalized tours in a museum and serendipitous items are introduced by slight diversions on the context-aware tours.

1 Background and motivation

Recommender Systems (RSs) try to assist users to access complex information spaces. They provide the users with personalized advices based on their needs, preferences and usage patterns. Moreover, common expectations concern with relevance, novelty and surprise. Various recommendation techniques have been investigated and each one has its own strengths and weaknesses. Especially, content-based techniques suffer of the *over-specialization* problem. Indeed, sometimes RSs can only recommend items that score highly against the user's profile and, consequently, the user is limited to obtain advices only about items too similar to those she already knows. Thus, the user can perceive the recommend items as obvious advices that are not so novel nor surprising. Indeed, novelty occurs when the system suggests an unknown item that the user might have autonomously discovered and a serendipitous recommendation helps the user to find a surprisingly interesting item that she might not have otherwise discovered (or it would have been really hard to discover) [5].

Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Piero Molino
Universitá degli Studi di Bari "Aldo Moro" - Dipartimento di Informatica, via E. Orabona 4, Bari (Italy), e-mail: {iaquinta, degemmis, lops, semeraro}@di.uniba.it, piero.molino@gmail.com

According to André et al. [1], the belief of serendipity as a valuable part of creativity, discovery and innovation is the main motivation of the interest of computer scientists about serendipity. Consequently, they have attempted to develop systems that deliberately induce serendipity and celebrated when it appeared as a side effect in systems built with other purposes in mind, for example the serendipitous discovery of something when browsing rather than searching hypertext documents [7]. However, most systems designed to induce or facilitate serendipity focus on the accidental nature of the serendipity and they neglect the breakthrough or discovery made by drawing an unexpected connection. Truly, the connections, though they may be guided, must remain unlooked for specifically to be considered serendipitous. Computer systems, however, may be able to help potential discoverers be as primed as possible to make unexpected connections in such a way that they are able to take advantage of them.

Our objective is to try to feed the user also with recommendations that could possibly be serendipitous. Thus, we propose to inject diversity in the recommendation task by exploiting the content-based user profile to spot potential surprising suggestions. In addition, the actual selection of serendipitous items is motivated by the real-world situation when a person visits a museum and, while she is walking around, she finds something completely new that she has never expected to find, that is definitely interesting for her. Thus, the applicative scenario concerns with personalized tours in a museum and serendipitous items are introduced by slight diversions on the context-aware tours. Indeed, the basic content-base recommender module allows to infer the most interesting items for the active user and, therefore, to arrange them according the spatial layout, the user behavior and the time constraint. The resulting tour potentially suffers from over-specialization and, consequently, some items can be found no so interesting for the user. Therefore the user starts to divert from the suggested path considering other items along the path with growing attention. On the other hand, also when the recommended items are actually interesting for the user, she does not move with blinkers, i.e. she does not stop from seeing artworks along the suggested path. These are accidental opportunities for serendipitous encounters. The serendipity-inducing module perturbs the optimal path with items that are programmatically supposed to be serendipitous for the active user.

The paper is organized as follows: Section 2 introduces the serendipity issue and covers strategies to provide serendipitous recommendations; Section 3 provides a description of our recommender system and how it discovers potentially serendipitous items in addition to content-based suggested ones; Section 4 provides the description of the experimental session carried out to evaluate the proposed ideas; finally, Section 5 draws conclusions and provides directions for future work.

2 Serendipitous recommendations

The idea of serendipity has a link with de Bono's "lateral thinking" [3] which consists not to think in a selective and sequential way, but accepting accidental aspects, that seem not to have relevance or simply are not sought for. This kind of behavior helps the awareness of serendipitous events, especially when the user is allowed to explore alternatives to satisfy her curiosity as in the museum scenario.

Moreover, serendipitous encounters depend on personal characteristics, e.g. the open minded attitude, the wide culture and the curiosity [9]. Therefore, the subjective nature of serendipity makes hard its conceptualization, its analysis and its implementation [4]. Anyway, programming for serendipity is feasible [2], for instance, by by allowing the users to expand their own knowledge and by preserving the opportunity of serendipitous discoveries.

Toms [10] suggests four strategies to introduce the serendipity: 1) Role of chance or 'blind luck', implemented via a random information node generator; 2) Pasteur principle ("chance favors the prepared mind"), implemented via a user profile; 3) Anomalies and exceptions, partially implemented via poor similarity measures; 4) Reasoning by analogy, whose implementation is currently unknown.

In [6] we propose an architecture for hybridizing a content-based RSs by the "Anomalies and exceptions" approach to provide serendipitous recommendations alongside classical ones. Thus, the basic idea underlying the proposed architecture is to ground the search for potentially "serendipitous" items on the similarity between the item descriptions and the user profile. More specifically, the problem of learning user profiles is managed as a binary *Text Categorization* task, since each document has to be classified as interesting or not with respect to the user preferences. Therefore, the set of categories is restricted to *POS*, that represents the positive class (user-likes), and *NEG* the negative one (user-dislikes). The content-based recommendations come of the matching of the concepts contained in the semantic profile and the concepts contained in the descriptions of items to be recommended. The recommended items are ranked according to the classification score against the *POS* and *NEG* classes. Thus, the list will contain on the top the most similar items to the user profile, i.e. the items high classification score for the class *POS*. On the other hand, the items for which the a-posteriori probability for the class *NEG* is higher, will ranked lower in the list. The items on which the system is more uncertain are the ones for which difference between the two classification scores for *POS* and *NEG* tends to zero. The uncertainty on the classification is used to spot items that are not known by the user, since the system was not able to clearly classify them as relevant or not.

3 Personalized museum tours

RSs traditionally provide a static ordered list of items according to the user assessed interests, but they are not aware about context facets concerning the user interaction with environment. Besides, if the suggested tour simply consists of the enumeration of ranked items, the path is too tortuous and with repetitive passages that make the user disoriented, especially under a time constraint. Fig. 1 shows a sample tour consisting of the *k* most interesting items, where the *k* value depends on how long should be the personalized tour, e.g., it deals with the overall time constraint and the user behavior. Moreover, different users interact with environment in different manner, e.g. they travel with different speed, they spend different time to admire artworks, they divert from the suggested tour. Consequently, the suggested personalized tour must be dynamically updated and optimized according to contextual information on user interaction with environment. The optimization task is performed by a genetic approach with a fitness function that relies on the user-sensitive time constraint, the user behavior (i.e., speed and stay times), the user learned preferences and the item layout.

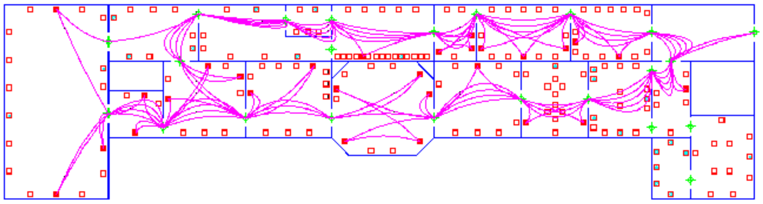


Fig. 1 A sample tour consisting of the ranked *k* most interesting items

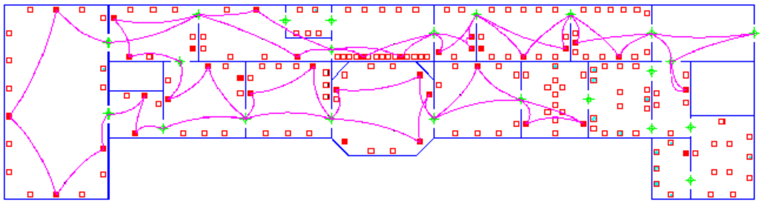


Fig. 2 Optimized version of the tour in Fig. 1

Once the personalized tour is achieved, as shown in Fig. 2, serendipitous disturbs are applied. The diversity injection is pursued by serendipitous disturbs to the the personalized tour. Indeed, the previous personalized tour is augmented with some items that are along the path and that are in the ranked list of serendipitous items according to the learned user profile and context facets [8]. The resulting path most likely has a worse fitness value and then a further optimization step is performed. However, the further optimization step should cut away exactly the disturbing serendipitous items, since they compete with items that are more similar with the user tastes. Therefore serendipitous items are differently weighed from the fitness function: their supposed stay time is changed. This implementation expedient also deals with the supposed serendipitous items should turn out not so serendipitous

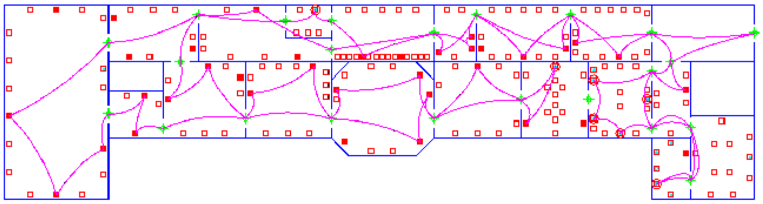


Fig. 3 The "good enough" augmented version

and the user should reduce the actual stay time in front of such items. Fig. 3 shows a "good enough" personalized tour consisting of the most interesting items and the most serendipitous ones. It is amazing to note that some selected serendipitous items are placed in rooms otherwise unvisited.

4 Experimental session

The goal of the experimental evaluation is to evaluate the serendipity augmenting effects on personalized tours. The dataset was collected from the official website of the Vatican picture-gallery and it consists of of 45 paintings and 30 users took part in the experiments.

The learned profiles were used to obtain personalized tours with different time constraints and different serendipitous disturbs. Five time (T_{10}, T_{15}, T_{20}, T_{25}, T_{30}) constraints were chosen so that tours consisted approximately of 10, 15, 20, 25, 30 items. Serendipitous items ranged from 0 to 7 (labels $S_0, \dots, S_7$).

The Table 1 reports the average of sums and means of *POS* values of tours. The serendipitous item augmenting causes the exploiting of items less similar to the user tastes according to her profile and this effect is particularly evident when there are too many serendipitous items. On the other hand, there is also a decrease when many items are selected according to the user profile, since they are progressively less interesting. When there are many items, the serendipitous item augmenting seems to have no effects over POS mean, but probably this comes from the not very large dataset used.

Table 2 reports percentages of walking time over the tour. Data show that, increasing the time constraint, less time is (relatively) spent to walk. Indeed, if few items are selected, they are scattered around (proportionally) many rooms and the user visits room with very few and even no one suggest item. The serendipitous item augmenting seems to increase the relative walking time. This result is quite amazing according to the selection serendipitous item strategy, i.e., items that are along to a previously optimized path. Actually, the walking time percentage mainly increases because serendipitous items are introduced as new genes of a "good enough" chro-

Table 1 Sums and means of POS values of tours

	T_{10}		T_{15}		T_{20}		T_{25}		T_{30}	
S_0	7.18	0.711	10.69	0.714	14.02	0.705	17.21	0.679	19.94	0.671
S_1	7.15	0.708	10.61	0.709	14.00	0.704	17.20	0.679	19.89	0.670
S_2	7.12	0.705	10.59	0.708	13.98	0.702	17.20	0.679	19.88	0.669
S_3	7.08	0.701	10.60	0.708	13.96	0.702	17.19	0.679	19.87	0.669
S_4	7.03	0.696	10.58	0.707	13.96	0.701	17.19	0.678	19.87	0.669
S_5	6.88	0.681	10.52	0.703	13.95	0.701	17.17	0.678	19.85	0.668
S_6	6.54	0.647	10.42	0.696	13.90	0.698	17.11	0.676	19.75	0.665
S_7	6.17	0.611	10.19	0.681	13.76	0.692	16.99	0.671	19.64	0.661
Items	10.10		14.97		19.90		25.33		29.70	

mosome (solution). However, the augmented chromosome tends to evolve toward the previous one. Thus the new genes should be promoted with a benefit over the fitness function: the reduction in their supposed stay time. This approach is simple and intuitive, but it makes difficult the interpretation of expected walking time percentage. Indeed, the variation on walking time becomes from path variations, but the total tour time is also changed on account of the technical issue about the genetic approach fitness function.

Table 2 Percentages of walking time

	T_{10}	T_{15}	T_{20}	T_{25}	T_{30}	
S_0	39.9	34.0	34.6	31.6	30.2	**34.1**
S_1	42.6	36.3	36.0	32.8	31.3	**35.8**
S_2	45.0	38.1	37.4	34.0	32.2	**37.4**
S_3	49.7	40.1	38.3	34.5	33.5	**39.2**
S_4	52.7	42.0	39.9	36.3	34.6	**41.1**
S_5	56.0	45.5	41.9	37.8	35.9	**43.4**
S_6	60.0	47.5	43.7	39.7	37.2	**45.6**
S_7	65.2	51.7	45.6	41.7	39.0	**48.6**

Moreover, the effects of serendipitous items on expected walking time are analyzed with respect to the starting optimized tours (S_0), i.e. the previously discussed drawback is partially cut off. Table 3 shows that few disturbs cause a quite uniform increase of the walking time percentage: the ground becomes from the slight deviations on S_0 tour. On the other hand, growing the number of serendipitous items, the deviations are amplified. This is more evident for the shortest S_0 tours, since many serendipitous items can encourage the "exploration" of rooms untouched by S_0, about Figure 3.

Table 3 Increment of walking time for tours with serendipitous items

	T_{10}	T_{15}	T_{20}	T_{25}	T_{30}
S_1	106	106	104	103	103
S_2	112	112	108	107	107
S_3	124	119	112	110	111
S_4	131	126	117	115	115
S_5	141	136	123	121	120
S_6	150	143	130	127	125
S_7	164	155	135	134	131

5 Conclusions and future work

This paper presents a beginning effort to apply some ideas about serendipity to information retrieval and information filtering systems, especially in RSs, to mitigate the over-specialization issue. Serendipity has a valuable part of creativity, discovery and innovation, but its subjective nature is problematic when trying to conceptualize, analyze and implement it. The attempts to develop systems that deliberately induce or facilitate serendipity often focus on the accidental nature of serendipity and the delight and surprise of something unexpected. On the other hand, they neglect the breakthrough or discovery made by drawing an unexpected connection. Thus, André et al. [1] stressed the importance of making use of serendipitous encounters in a productive way.

Hence, the evaluation of recommendations has to be further investigated. Indeed, the recommendation process relies on the provided ratings and they should be also interpreted according to the serendipity point of view. Used ratings are often too synthetic and, consequently, they conceal the user rating motivations that affect the meaning evaluation of finding unknown and possibly interesting things, and not simply interesting ones. For instance, a poorly rating for suggested items should come from the experience of the user (the user already knows the item), from her lack of interest (the user already knows the item and is not interested in it), from her lack of interest in finding new things (the user does not know the item and has no interest in knowing something new), from the conscious expression of dislike (the user did not know the item before, now she knows it but she does not like it or is not interested in it) or from a serendipitous encounter (before-unknown item that results to be interesting for the user).

The museum scenario is particularly interesting because items are arranged in a physical space and users interact with the environment. Thus disregarding context facets makes useless recommendations.

Similar remarks are still valid in domains (different from cultural heritage fruition) in witch a physical or virtual space is involved and it represents a pragmatic justification to explain (supposed) serendipitous recommendations. Item descriptions are the starting point to exploit content-based methods to implement a hybrid RS that is aware of contextual facets and that uses them, in concurrence with semantic profiles, to spot serendipitous items.

As future work, we expect to carry out more extensive experimentation with more users and wider item collections. We plan also to gather user feedback and feeling by questionnaires focused on qualitative evaluation of the recommendations and the idea of getting suggestions that should surprise them. That is really important for the need to understand the effectiveness of the module in finding unknown items rather the ones that result best rated. Experimentation with users with different cultural levels and with different information seeking tasks are also important to find out which kind of user would like most serendipitous recommendations and to whom they are more useful.

References

1. André, P., Schraefel, m., Teevan, J., Dumais, S.T.: Discovery is never by chance: designing for (un)serendipity. In: Proceeding of the seventh ACM conference on Creativity and cognition (C&C '09), pp. 305–314. ACM (2009)
2. Campos, J., de Figueiredo, A.: Searching the unsearchable: Inducing serendipitous insights. In: R. Weber, C. Gresse (eds.) Proceedings of the Workshop Program at the 4th International Conference on Case-Based Reasoning (ICCBR 2001), pp. 159–164 (2001)
3. De Bono, E.: Lateral Thinking: A Textbook of Creativity. Penguin Books, London (1990)
4. Foster, A., Ford, N.: Serendipity and information seeking: an empirical study. Journal of Documentation **59**(3), 321–340 (2003)
5. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
6. Iaquinta, L., de Gemmis, M., Lops, P., Semeraro, G., Molino, P.: Can a recommender system induce serendipitous encounters? In: K. Kang (ed.) E-Commerce, pp. 227–243. In-Teh (2010)
7. Marchionini, G., Shneiderman, B.: Finding facts vs. browsing knowledge in hypertext systems. Computer **21**(1), 70–80 (1988)
8. Mehta, B., Niederée, C., Stewart, A., Degemmis, M., Lops, P., Semeraro, G.: Ontologically-enriched unified user modeling for cross-system personalization. In: L. Ardissono, P. Brna, A. Mitrovic (eds.) Proc. of 10th Int. Conf. on User Modeling (UM2005), pp. 119–123. Springer (2005)
9. Roberts, R.M.: Serendipity: Accidental Discoveries in Science. John Wiley & Sons, New York (1989)
10. Toms, E.G.: Serendipitous information retrieval. In: DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries. Zurich (2000)

RefGen: Identifying Reference Chains to Detect Topics

Laurence Longo, Amalia Todirașcu

LiLPa laboratory, University of Strasbourg, 67000 Strasbourg, France

Abstract. In this paper, we present *RefGen,* a reference chain identification module for French. *RefGen* is part of a topic detection system, used to improve a search engine by topic indexing. *RefGen* algorithm uses genre specific properties of reference chains and (Ariel 1990)'s accessibility theory. It applies strong and weak filters (lexical, morphosyntactic and semantic filters) to automatically identify coreference relations between referential expressions. We evaluate the results obtained by *RefGen* from a public reports corpus.

1. Introduction

We present a project aiming at automatic topic detection, by combining statistical methods and linguistic methods. We use several linguistic cues to detect topic changes: discourse markers, reference chains and theme/rheme positions. In this paper, we focus on the reference chain identification module *RefGen*, one of the main modules of our topic detection system, integrated into a topic search engine. The search engine uses topic indexing to help users to retrieve relevant documents from the archives.

Beside the use of explicit discourse cue phrases, we assume that topics will be mainly discovered from linguistic markers as reference chains and anaphora pairs (Cornish 1995), (Schnedecker 1997). After the identification of reference chains, we propose local candidate topics for each document segment, selecting them from the first elements of the reference chains. While our goal is to select a global topic associated to the document, we select it from these local topics, by applying selection criteria as frequency (several chains referring to the same entity), as position in the document, as topic continuity among several segments (Goutsos 1997).

Identifying reference chains is a key process for many NLP applications as topic detection, text summarisation. To solve the reference, the systems identify the various referential expressions (e.g. pronouns, definite noun phrase, possessives) referring the same discourse entity. This is a criterion to form a reference chain in the document. A reference chain includes at least three reference expressions (e.g. *Barack Obama... il... lui*) which denote the same referent (Schnedecker 1997). This referent is common to several sentences of the same paragraph and it represents a potential topic candidate. Coreference resolution methods either apply heuristic rules manually defined (which select the most suitable candidates for a given anaphor) or rules learned from annotated corpora. While supervised learning methods (Ng and Cardie 2002), (Hoste 2005) are effective in the processing of coreference relations, they require large, manually annotated training corpora. However, there is currently no large reference corpus annotated with reference chains in French[1] (Salmon-Alt 2001) that might be used to apply machine learning techniques.

To identify reference chain expressions, we propose a new knowledge poor method as adopted for pronoun (Mitkov 1998) and coreference resolution (Hartrumpf 2001), (Popescu-Belis 1999), (Bontcheva et *al*. 2002). We select the coreference chain elements using criteria about accessibility and information content of various categories of referring expressions (Accessibility theory (Ariel 1990)), their syntactic function, but also some genre-dependent properties of reference chains. The *RefGen* algorithm proceeds in two steps: it first selects the starting element of a reference chain and then it selects the next elements of the reference chain from a list of antecedent-anaphor potential pairs. These pairs verify strong and weak constraints (lexical, syntactic and semantic) (Gegg-Harrisson and Byron 2004) between antecedent-anaphor potential pairs.

The paper is organised as follows. In section 2 we present the archi-

[1] For example, SemEval 2010 task#1 *Coreference Resolution in Multiple Languages* campaign provides training data for different languages except French.

tecture of our topic detection system. In section 3 we describe the *RefGen* module: the genre-dependent parameters used to identify chains and the corpus analysis, the annotation module and the algorithm. We then discuss the *RefGen* results obtained from a comparison with manually annotated corpora. In section 4 we conclude and we present future developments.

2. The Architecture of the Topic Detection System

For our project, we consider that the topics are aggregates of the sentence themes (Goutsos 1997), while sentence themes are actors, ideas or events. To detect topics, we use the global properties of the text: cohesion and coherence (Halliday and Hasan 1976), but also genre-specific properties (Biber 1994). Thus, we combine statistical methods (Choi et al. 2001) and linguistic markers identification. Beside the use of explicit discourse cue phrases (Charolles 1997), we assume that topics will be mainly discovered from cohesion markers, as reference chains (Schnedecker 1997) and anaphora pairs (Kleiber 1994).

We present our topic detection system's architecture, which is still under development. First, we convert the documents available in various formats (PDF, XML, etc.) to raw text. Then, we segment the documents into several homogeneous units, using C99 algorithm (Choi et al. 2001). This module detects the boundaries of the topic homogeneous units using lexical-based cohesion measures, but it does not explicitly extract topics from the unit. To associate topic candidates to each segment, we apply several heuristic rules exploiting linguistic information. Thus, we use two categories of linguistic markers: explicit discourse cues, used to focus on a specific topic (Charolles 1997), (Porhiel 2004) and cohesion markers (as reference chains (Schnedecker 1997) and anaphora pairs (Kleiber 1994)). We establish a list of discourse markers (as *concernant X, au sujet de X..., dans un premier temps, finalement*) explicitly indicating the topic of the sentence or of the paragraph. Reference chains or anaphora pairs indicate that the same entity is referred several times in the text, the introduction of a new entity and of a new reference chain is a sign of topic shift. The core module of our topic detection system is *RefGen*, the reference chain identification module. This module uses several lexical, syntactic and semantic constraints to detect the referring expressions and it will be described in section 4.

The output of the topic detection module is a set of topics associated to each document. We apply the reference chain identification algorithm to each segment, to propose some local topic candidates. We select the starting elements of the reference chains as local topics. Then, we check criteria as frequency (several chains referring to the same entity), as position in the document (if the topic occurs in the title or in the first paragraph of the document), as topic continuity among several segments, to propose topic candidates describing the document.

The topics extracted from each document are used by the search engine to index the document. Thus, it is possible to select associated documents, among the documents indexed by similar topics.

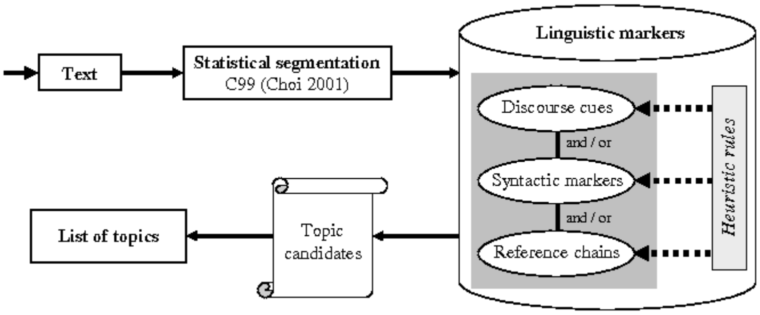


Fig. 2.1 The architecture of the topic detection system

3. The Reference Chains

Following (Schnedecker 1997), we consider a reference chain as a relation between at least three mentions (three referential expressions). The reference chains include three types of constituents with a referential function: the proper names, the NPs (definite, indefinite, possessive or demonstrative) and the pronouns. The proper names play an important role in the discourse structure due to the fact that they are often opening a reference chain in the journalistic portraits (Schnedecker 2005). Apart from cases where there is a referential competition (the repetition of the proper name eliminates ambiguity between two referents), the repetition of a proper name signals a break in the reference chain. When a referring expression is used, it triggers a "particular recruitment process" of a referent. Thus, the demonstrative (e.g. *ce président*) points directly to the referent on the basis of proximity while the anaphoric pronoun "*il*" recruits a referent that is the argument of a salient phrase (Kleiber 1994). The use of a particular mention (referential expression) is an indication for the reader to remember a specific referent and which is a local theme. However, the use of noun phrases in contexts that a pronoun would be sufficient is an indication of a reference break. These informations will be used by the topic detection system.

We process single referential relations (excluding plural anaphora) between co-referent expressions, between and within sentences. We treat direct coreference (Manuelian 2002) for the coreferential NPs having the same head (eg "*le changement climatique / ce changement*") and some indirect coreference between Person and function name (e.g. "*Barack Obama … le président*").

4. The *RefGen* Module

We assume that reference chains have specific linguistic properties depending on the text genre and type (explanatory, narrative etc.) and we exploit these properties for reference chains identification.

In the section below, we present the study of reference chains properties on a corpus of several genres. Then, we focus on the *RefGen* module and on the linguistic annotations required (tagging, chunking and Named Entities recognition). We explain the reference chains algorithm (*calcRef*) before presenting the results of the evaluation.

4.1. Corpus Analysis

To identify reference chains genre specificities we study the reference chains in a French corpus (about 50 000 tokens) composed of five genres: newspapers from *Le Monde* (2004), editorials from *Le Monde Diplomatique* (1980-1988), a novel *Les trois Mousquetaires* (Dumas, 1884), some European legal standards from the *Acquis Communautaire* (Steinberger et al. 2006) and public reports from *La Documentation Française* (2001) (Longo and Todirascu 2010). We manually annotate the chains to determine which reference chain properties are relevant for a particular genre.

The reference chains study is based on (Schnedecker 2005). For each genre, we examine the chains following five criteria:

- the average length of chains (the number of referential expressions);
- the average distance between the elements of a chain (the number of sentences);
- the frequency of the elements depending on their grammatical class;
- the grammatical class of the starting element of a chain;
- the identity between the sentence theme and the first element of a chain.

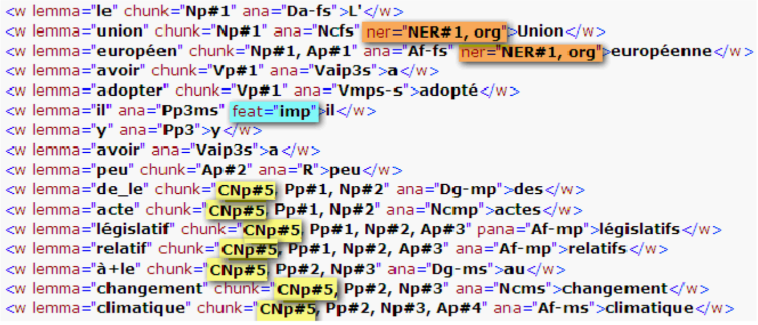
The study reveals several differences. For example, the average length of reference chains from *Acquis Communautaire* is three, while the length is nine for *Les trois Mousquetaires*. Concerning the frequency of the reference element classes, we notice that *le Monde* contains mostly Proper Nouns (30.8 %) while *Le Monde Diplomatique* contains 50 % of definite NP. Proper Nouns are very frequent starting elements for newspapers reference chains, but indefinite NP for *Acquis Communautaire*. In addition, the first element of the chain is the sentence theme for 80 % of the occurrences for the newspapers and only for 40 % for the public reports.

Thus, the corpus analysis on the reference chains highlights their genre-specific properties. We use these parameters to configure *RefGen* according to the genre.

4.2. Annotation of the Referential Expressions

To identify the referential expressions we tagged, lemmatized and chunked the documents using TTL tagger (Ion 2007). This tagger identifies chunks (simple noun phrases, prepositional phrases) and morpho-syntactic properties (tense, mode, person, gender, number). Then, we apply a set of rules to identify complex NP (NP modified by at most two PP, NP modified by a relative clause), as "*l'élévation du niveau global de la mer*" which are more informative than simple NP. While we search topic entities, we apply some heuristic rules to identify persons and organizations. In addition, we annotate the French impersonal pronoun *il* (e.g. "*il* pleut", *it* rains) to eliminate the non-anaphoric use of this pronoun.

Fig.4.2.1 is an example of annotations including lemmas (`lemma`), chunks (simple `Np`, `Pp`; complex `CNp`), morpho-syntactic properties (`ana`), named entities (`ner`) and impersonal pronoun *il* (`feat="imp"`). We use these linguistic annotations to identify the reference chains and anaphora pairs.



```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma="il" ana="Pp3ms" feat="imp">il</w>
<w lemma="y" ana="Pp3">y </w>
<w lemma="avoir" ana="Vaip3s">a </w>
<w lemma="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" pana="Af-mp">législatifs</w>
<w lemma="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Fig. 4.2.1 Example of annotated output in *RefGen*

4.3. The CalcRef Module

CalcRef is the main module of *RefGen* and it proceeds to reference chain identification by using genre-depedent parameters and the linguistic annotations presented in the previous section. Thus, we specify the genre of the indexed documents and *CalCRef* is configured according to the properties of the reference chains (average distance between the mentions, average chain length, the preferred category of the first element of a chain). Thus, for a corpus of public reports, we use the length of 4, the average distance is 2 phrases and the preferred type of the first element is a complete definite NP.

For each paragraph, *CalcRef* selects candidates for the first mention of reference chains, among expressions with a high degree of information content. (Ariel 1990) defines an accessibility hierarchy to classify the referential expressions according to their accessibility: less the referent is accessible, the referential expression should be longer. Thus, indefinite NPs, Proper Nouns or complete NP (definite NP modified by PPs or by a relative clause), occupying the thematic position are used to mention a new entity, while short mentions[2] as pronouns might be used to refer to entities already specified in the discourse. The accessibility is computed by combining three elements: informativity (the amount of lexical information), rigidity (the possibility to pick up a specific referent) and attenuation (phonological size). We propose weights on a scale of 10 to 110 for each element (the global weight of the complete proper noun "*Le president Barack Obama*" is 220 while it is 150 for the pronoun "*elle*").

CalcRef computes the global weight of the candidates as a sum of the global accessibility weight and the syntactic role weight. We also define a scale for the syntactic role weights: 100 for the subject position, 50 for the direct object position, 30 for the indirect object and 20 for other syntactic functions. Then, genre dependent parameters (the preference for the first element type or the distance between the mentions) are used to increase the weight (+50) of some candidates (for example, if we treat law texts, indefinite NPs are preferred as starting elements a chain). We order the first element candidates according to the global weight and we select the highest weight candidate as a first element of the chain.

CalcRef selects the next elements of the reference chain from highly accessible expressions (pronouns, demonstratives etc.). We identify

[2] Mentions means also referential expression

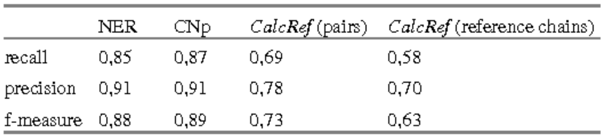
potential antecedent-anaphor pairs, if the distance between the two elements is less than the average distance defined by the genre parameters. Then, we adapt the method proposed by (Gegg-Harrison and Byron 2004) applying several constraints between antecedent and anaphor to filter out impossible pairs. For each pair, we check some strong and weak constraints. Weak constraints mean that they might not be satisfied, even if there is a valid antecedent-anaphor pair: agreement in gender or number, similar syntactic function, semantic knowledge (Persons might be valid antecedents of a NP expressing a function). Strong constraints concern imbrication (an element must not be nested in its antecedent as *[la soeur [de Marie]]*), co-arguments of a verb should not be coreferent etc. For each candidate pair satisfying the strong constraints, we check the number of the weak constraints that are satisfied. In the case of the several pairs satisfying the same number of constrains for the same anaphor, we keep the valid pairs into a large list.

Then, we start from the first element of the chain and we search the pairs having this candidate as antecedent in the list. To build the reference chain, we apply the transitive property: if A is antecedent of B and B is antecedent of C, then they are part of the same chain. We continue the process until the length of the current reference chain is greater than the average gender-specific length. We annotate the candidate pairs identified as part of the current reference chain.

We restart the whole process after selecting the next first candidate element from the ordered list.

4.4. Evaluation

We present the first results of evaluation of *RefGen*, we compare the reference chains extracted automatically against the manually annotated corpus. We present the results obtained for the CNp annotation module, for the NER module and for the chain identification module. The evaluation corpora is a small corpus (7230 tokens) composed of public reports of the European Commission about the climate changes and the measures adopted by EU to limit the effects of the climate changes. We compute the recall, the precision and the f-measure of the intermediate modules, as well as the results for *CalcRef*. We check the results obtained for independent antecedent – anaphor pairs, as well as for reference chains.



	NER	CNp	*CalcRef* (pairs)	*CalcRef* (reference chains)
recall	0,85	0,87	0,69	0,58
precision	0,91	0,91	0,78	0,70
f-measure	0,88	0,89	0,73	0,63

Table 4.4.1 The evaluation of *RefGen*

The NER annotation errors (from the PR corpus) are due to the wrong identification of some acronyms or abbreviations (e.g. *GES: gaz à effet de serre*) which were annotated as Organizations. The CNp identification module fails to identify some CNps (an NP modified by more than three PP), which were not described by the existing set of patterns. Indeed, the test corpus is characterized by very frequent, complex, informative noun phrases. In contrast, the newspaper corpus is rich in NE. For the PR corpus, *CalcRef* identifies 118 pairs, but only 24 could be related by reference chains. Several antecedent-anaphor pairs were wrongly selected, due to tagging errors or due to the insufficient external knowledge sources. For example, some of the antecedent-anaphor pairs were selected because they satisfy the same number of constraints (number, gender, syntactic function).

If we change the genre parameters (distance between referential expressions, length, type of the first element of a chain), we obtain quite similar results for the pairs (*f_measure* is 0,70) and worse results for the reference chains (*f_measure* is 0,54).

5. Conclusion

We presented *RefGen*, a reference chain identification module, developed for French. This module uses a set of detailed linguistic annotation and the accessibility hierarchy of the referring expressions to select possible antecedent-anaphor pairs. Then, a set of lexical, syntactic and semantic constraints are used to filter some pairs. *RefGen* also uses some genre-dependent properties of the reference chains (average length, preferred type of the first element, average distance separating several mentions of the same referent). These genre-dependent properties were identified from a corpus-based analysis. We describe the algorithm adopted to identify the reference chains and we present a first evaluation of the module. In the future, the module will be integrated into the topic detection system. Future work focus on the adaptation of the system for other languages.

6. References

Ariel M. (1990). *Accessing Noun-Phrase Antecedents*, London: Routledge.

Biber D. (1994) Representativeness in corpus design. *Linguistica Computazionale*, IX-X, Current Issues in Computational Linguistics: in honor of Don Walker.

Bontcheva K., Dimitrov M., Maynard D., Tablan V., Cunningham H. (2002). Shallow methods for named entity coreference resolution. *Proceedings of TALN 2002.*

Cornish F. (1995). Références anaphoriques, références déictiques, et contexte prédicatif et énonciatif. *Sémiotiques*, 8, 31-57.

Gegg-Harrison W., Byron D. (2004). PYCOT: An Optimality Theory-based Pronoun Resolution Toolkit. *Proceedings of LREC 2004*, Lisbonne.

Goutsos D. (1997). *Modeling Discourse Topic: sequential relations and strategies in expository text.* Norwood, N.J.: Ablex Publishing Corporation.

Hartrumpf S. (2001). Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics. *Proceedings of CoNLL (Computational Natural Language Learning Workshop).*

Hoste V. (2005). Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, 246 p.

Ion R. (2007). *TTL: A portable framework for tokenization, tagging and lemmatization of large corpora.* Bucharest : Romanian Academy.

Kleiber G. (1994). *Anaphores et Pronoms.* Louvain-la-Neuve : Duculot.

Longo L., Todirascu A. (2010). Une étude de corpus pour la détection automatique de thèmes. *Proceedings of the 6th journées de linguistique de corpus (JLC 09*), Lorient.

Mitkov R. (2001). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence: An International Journal*, 15, 253-276.

Ng V., Cardie C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the ACL (Association For Computational Linguistics),* Morristown, 104-111.

Popescu-Belis, A. 1999. Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation. Thèse d'Université, Université Paris-XI.

Porhiel S. (2004) Les introducteurs thématiques, *Cahiers de Lexicologie*, 85

Salmon-Alt S. (2001). Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel. PhD thesis, Université H. Poincaré, Nancy.

Schnedecker C. (1997). Nom propre et chaînes de référence. *Recherches Linguistiques* 21. Paris : Klincksieck.

Schnedecker C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de Linguistique* 51, 85-133. Duculot.

Reasoning in a Distributed Semantic Indexing System

Claude Moulin[1], **Cristian Lai**[2]

[1]University of Technology, Heudiasyc CNRS,

Compiègne, France

claude.moulin@utc.fr

[2]CRS4, Center of Advanced Studies, Research and Development in Sardinia,

Parco Scientifico e Tecnologico, Ed. 1

09010 Pula (CA), Italy

clai@crs4.it

Abstract This paper focuses on the semantic indexing of resources in a peer to peer network. Keys used for indexing and the corresponding urls of indexed resources are stored in a distributed hash table, scattered on the different peers of a community. A key is a semantic description of resources and can be considered as a small knowledge base because it refers to concepts and properties belonging to ontologies. In this kind of index a key used for resource retrieval must be identical to a key used for indexing the resource. Therefore, it is necessary to publish a resource with different keys allowing its further retrieval. We have defined an expansion mechanism of the small knowledge base that constitutes the semantic description of a resource. With different examples, we present the main cases of expansion that define the retrieval context of a resource.

1 Introduction

In the world of communities, people are strongly motivated in sharing resources with respect to their interest. In our research, we mainly deal with communities whose members are scattered in a distributed Peer to Peer (P2P) network. In this case, people are not interested in maintaining direct contacts and exchanging messages, or having activities typically facing centralized approaches like in current social Web platforms.

Resources interesting for the members of such communities can have different types like documents or notes and can have different formats like text, image or video. The main activities that community members do regularly, on one hand, concern the indexing of own resources and their publication in the P2P network and, on the other hand, the resources retrieval from the same network. The community nature is not relevant because the solution we propose is not specific to a domain. However, most of the examples we give belong to the e-learning domain.

The main problem raised by this situation is twofold: (i) what kind of system is better for indexing resources knowing that communities' members are not specialist of software installation and don't want to have a complex system to maintain; (ii) how to retrieve a document thanks to an index which is distributed on different peer computers within a response time as short as possible. In such a Boolean index (Salton et al., 1982), keys used for resources retrieval must be equals to keys used for publication. The concomitant problem, which is not specific of this approach but is more crucial in these circumstances, is the difference of context between the publication and the retrieval of a resource.

In this paper we propose to develop some aspects of the solution we brought to this problem. We have chosen a semantic indexing of the resources based on ontologies. The choice of the ontologies is free and these documents are also published in the network. All the keys used for representing a resource in the index embody its semantic description and are written in a language based on RDF. A resource may be indexed by more than one key and a key may index several resources. We can consider the index as a distributed knowledge base. We fasten to a document the meta-information that describes its content and we publish it in the network. The URL of the resource is tied up to the meta-information and allows its further access.

The problem to solve is the difference between a publication context and a retrieval context. As a solution, we propose to foresee during the publication of a resource, different reasonable retrieval situations, and therefore different queries to which the resource should respond positively. We use a reasoning based on the ontologies involved in the semantic description of a resource.

For describing the different examples presented in this paper, we use the following set of ontologies: lom.owl (denoted by *lom*) (Ghebghoub, 2008, 2009), describes the domain of learning objects; lt.owl (denoted by *p2p-lt*) describes the concepts of the theory of languages; system.owl (denoted by *system*) is an ontology we have developed for representing the resources of our system and simplify an important, maybe the most important indexing case.

2 Reasoning

A reasoner is a piece of software able to infer logical consequences from a set of axioms or asserted facts. In practice a reasoner makes inferences either about classes constituting an ontology or individual constituting a knowledge base. Ontology classification arranges classes defined by logical expressions into a hierarchy. This reasoning task is normally related to ontology development.

Our approach concerns the query answering with respect to ontology based information retrieval. The Semantic Web requires high-performance storage and reasoning infrastructure in order to match the demand of indexing structured data with the use of ontologies. The major challenge toward building such infrastructure is the expressivity of the underlying standards such as RDF(s) and OWL (Kiryakov et al, 2005).

For the first case of reasoning (concept classification), we may use different available engines related to ontology languages for the Semantic Web like OWL and RDF(s) (Jing Mei et Parsia, 2004). Among the most popular there are RacerPro, FaCT++ and Pellet. In particular, Pellet is an OWL DL reasoner based on the tableaux algorithm (Baader, 2001) developed for expressive Description Logics. Pellet parses OWL documents into triples and separates them into TBox (axioms about classes), ABox (assertions about individuals) and RBox (axioms about properties), which are passed to the tableaux based engine. Logic relations contained into the ontology and constituting classes, individuals, properties allows to create new axioms. An interesting feature of Pellet is its strictly relation to ontology analysis and repair. In fact, as explained in (Parsia et Sirin, 2004) OWL has two major dialects, OWL DL and OWL Full, with OWL DL being a subset of OWL Full. All OWL knowledge bases are encoded as RDF/XML graphs. OWL DL imposes a number of restrictions on RDF graphs, some of which are substantial (e.g., that the set of class names and individual names be disjoint) and some less so (that every item have a type triple). Ensuring that an RDF/XML document meets all the restrictions is a relatively difficult task for authors, and many existing OWL documents are nominally OWL Full, even though their authors intended for them to be OWL DL. Pellet incorporates a number of heuristics to detect DLizable OWL Full documents and repair them, i.e. making them compliant with DL characteristics.

In our work we don't take care of the design of ontologies but we assume to find well structured and ready to use ontologies. We are not interested in ontology analysis and repairing. However, manual semantic indexing is only possible with named concepts with clear descriptions. We only propose this kind of concepts in our tools for building semantic descriptions. It is very difficult to deal with anonymous concepts built on logical expressions.

The second case of reasoning based on the structure of the ontology applies the semantics rules of OWL to a knowledge base. This adds new assertions to the knowledge base (the first case adds new axioms to the ontology). In (Kiryakov et al, 2005) the authors explain two principle strategies for rule-based inference, forward-chaining and backward-chaining. Their approach is based on inferred closure and known as materialization. Through the inferred closure, a knowledge base is extended with all the facts inferred by the application of semantic rules.

In our case, the system builds a semantic description of a resource. This description is a small knowledge base that has the form of a tree of nodes. The root element represents the resource to index. Other nodes are ontology identified elements or anonymous local resources allowing to create a path from the root node (see Fig 1). We also need an expansion mechanism inferring from only a few semantic rules, mainly relative to subsomption. The application of other semantic rules would infer facts about anonymous elements contained in the small and local knowledge base. It is useless because it is not concerning the resource to be indexed. Our work intends to generate materialization within a virtual knowledge base created each time for a specific case.

3 Semantic indexing

We produce two examples for defining the notion of publication and retrieval contexts.

3.1 Publication and Retrieval Contexts

Let's consider an example related to the ontology denoted by *lom*, concerning the learning object domain. We can say that a resource has for *difficulty* level, *very difficult*:

```
[a lom:LearningObject] lom:hasLomEducational [lom:hasDifficulty lom:veryDifficult]
```

We are speaking about a resource whose *lom:LomEducationaCategory* has for *lom:difficulty lom:veryDifficult*. We create a virtual knowledge base with two instances *_:lo* and *_:lec* (for clearness we use the N3 Notation). *_:lo* is an instance of the class *lom:LearningObject*, subject of the relation *lom:hasLomEducational* with the object *_:lec*, that should be specified.

```
_:lo
   rdf:type     lom:LearningObject ;
   lom:hasLomEducational _:lec .
```

_:lec gives the level of difficulty and is subject of the relation *lom:LomEducationalCategory* with lom:very_difficult, an instance of the class *lom:Difficulty*.

```
_:lec
   a     lom:LomEducationalCategory ;
   lom:hasDifficulty lom:very_difficult .
```

RDF's conceptual model is a graph (Beckett et Berners-Lee,2008). Fig 1 shows the graphical representation of the small knowledge base representing the *_:lo* resource description.

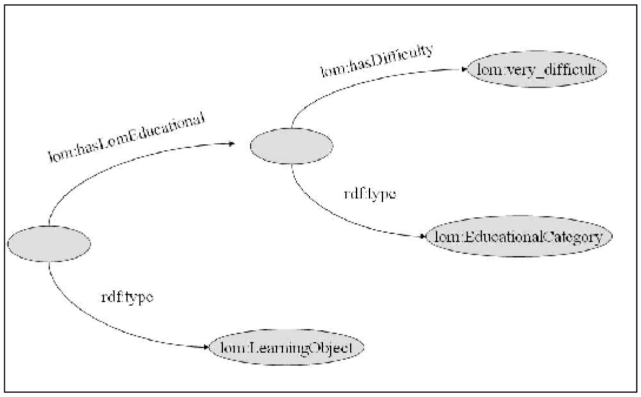


Fig 1. The RDF graph of LearningObject.

We define the publication context of the *_:lo* resource with the following triples:

```
_:lo
    rdf:type        lom:LearningObject ;
    lom:hasLomEducational _:lec .
_:lec
    lom:hasDifficulty lom:very_difficult .
```

The resource is identified by its type and the composition of properties where it is involved. A description corresponds to a tree of RDF triples containing a path, i.e. a sequence where the object of a triple is the subject of the following one. The other triples denote the link between an instance and a concept. The publication context is used to create the key associated to the resource in the distributed index. The anonymous nodes identifiers are cancelled.

```
Key_1:
{rdf:type,lom:LearningObject,lom:hasLomEducational,lom:hasDifficulty,lom:very_difficult}
```

For resource retrieval, a key must be supplied and must be equal to the key created from the publication context. A retrieval context is the description of a required resource and must correspond to a publication context; the identifiers of the anonymous nodes do not matter. In the example, Key_1 should be created. However, it is interesting that the resource (denoted by *_:lo*) can be found from other queries. For example, a user may be interested by learning object where the difficulty level has been defined, no matter which level. In this case, the retrieval context is a graph pattern containing a variable for representing the undefined value.

```
_:lo
    rdf:type      lom:LearningObject ;
    lom:hasLomEducational  _:lec .
_:lec
    lom:hasDifficulty ?x .
```

In this case the retrieval key is:

```
Key_2:
{rdf:type,lom:LearningObject,lom:hasLomEducational,lom:hasDifficulty}
```

In the same way, we have identified different situations where we define several retrieval contexts from the same publication context. This definition requires an expansion mechanism which is presented in section 4.

Generally, semantic indexing is the attachment of a resource to a concept. In this case, the provider means that the subject of the resource is about a concept defined in an ontology but without any other specification. We have considered this eventuality and created a system ontology with the concept of document and the relation of interest for representing it

In the following example, the resource *_:d* is about the concept of Stack (extracted from the ontology on the theory of language denoted by p2p-lt). For maintaining a DL ontology, we cannot associate a resource to a concept. We consider that the resource concerns any representative instance of the concept.

```
_:d
    a      system:Document ;
    system:hasInterest _s.
_:s rdf:type p2p-lt:Stack .
```

The concept of Stack has a super-concept: DataStructure. Fig 2 shows the graphical representation of the small knowledge base representing the *_:d* resource description.

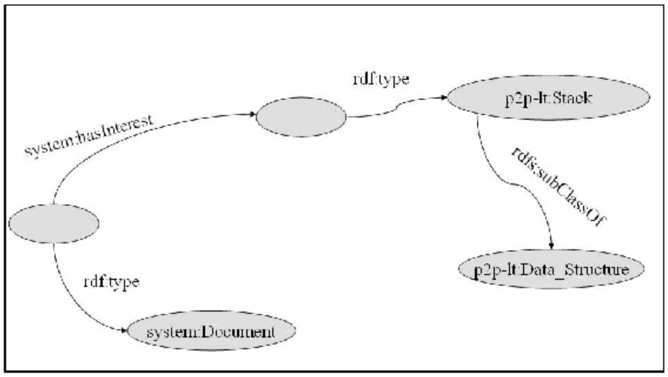


Fig. 2 . The RDF graph of Stack.

In order to expand the publication context of the *_:d* resource, we have to consider the subsumption on the Stack concept. _:s has type Stack and then has also the DataStructure type. We can identify two keys:

```
Key_1:
    {system:hasInterest,rdf:type,p2p-lt:Stack}
Key_2:
    {system:hasInterest,rdf:type,p2p-lt:Data_Structure}
```

Then we reduce them to:

```
Key_1:
    {system:hasInterest,p2p-lt:Stack}
Key_2:
    {system:hasInterest,p2p-lt:Data_Structure}
```

Each resource may be considered as an instance of a concept of one or more ontology. It is possible to combine several elementary descriptions of resources and to create as many keys.

3.2 How to Find Ontologies for Indexing?

The ontologies used for the resource descriptions are published in the P2P network as the other resources. A special key using the system ontology allows their publication. No expansion is necessary. We provide for a tool that helps the users to navigate the ontologies and automatically create the keys corresponding to the publication and retrieval contexts. However, the process of ontology publication must start else the discovery would be impossible.

We consider that some expert users have the skills to look for and select ontologies interesting the community members (Gruber, 1993). Our system allows the publication each time a new ontology is useful for the community and can be shared. It also requires a small description and the application domain of this ontology. It is not possible to cancel an ontology. Most of the users are not aware of the existence of ontology and are not involved in this process.

When there is no ontology for describing with precision a resource, it is ever possible to associate a resource with a concept thanks to the system ontology.

4. Concept of Expansion

A description represents a context of use, i.e. the conditions and circumstances that are relevant to an event of publication and retrieval. To overcome the difference between a publication context and a retrieval context it is necessary to foresee during the publication of a resource, different reasonable retrieval situations, and therefore different queries to which the resource should respond positively.

The general mechanism is the following. The resource provider builds the semantic descriptions of the resource which constitute the publication context. Each description is expanded by one or more descriptions that constitute the retrieval context. During the publication, all the descriptions belonging to the publication and the retrieval contexts are translated in keys and the resource is indexed by each of these keys. We consider that a retrieval context is either a generalization of a publication context or a close context. In consequence, we expand a publication context with descriptions corresponding to the most likely retrieval contexts. A resource is then published with different keys generated from the one created by the resource provider. Even if the time required for publishing a resource is a little bit longer, at this stage it is not penalizing.

There are many cases we have considered with respect to the expansion. As a very simple example, the following figure shows that the generalization of a concept in one ontology leads to an expansion of a context (using the Manchester syntax[1]). If a document concerning the concept of Stack is published, it should be retrieve from a query on documents talking about data structures.

[1] http://www.w3.org/TR/owl2-manchester-syntax/

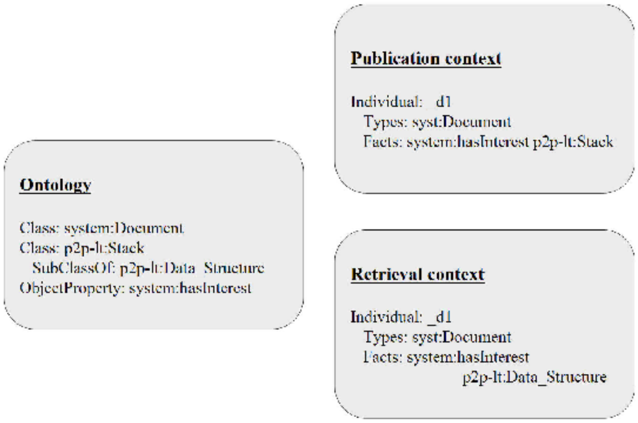
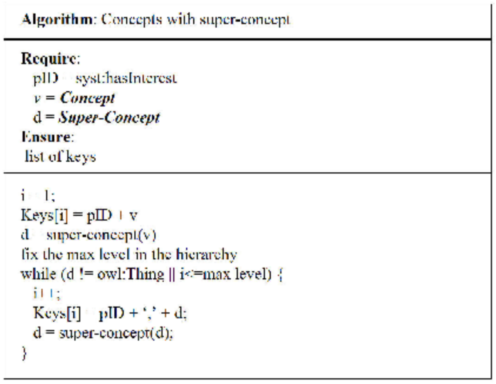


Fig. 3. Generalization of a concept.

The ontology level shows the generalization of concept Stack. The key of publication are created thanks to an algorithm whose pseudo-code is the following:



```
Algorithm: Concepts with super-concept

Require:
  pID   syst:hasInterest
  v = Concept
  d = Super-Concept
Ensure:
 list of keys

i   1;
Keys[i] = pID + v
d   super-concept(v)
fix the max level in the hierarchy
while (d != owl:Thing || i<=max level) {
  i++;
  Keys[i]   pID + ',' + d;
  d = super-concept(d);
}
```

Fig. 4. Pseudo-code for key generation algorithm.

Key are under the shape {pID, v}, where pID is the property system:hasInterest and v is the concept. A loop determines the super-concepts within the hierarchy of the ontology. A maximum number of iterations are fixed in order to limit the generalization level.

Our model of expansion associates the publication context and a sub-graph of the ontology concerned by this context in order to generate a retrieval context. A specific algorithm allows generating it.

We have classified all the expansions. The path contained in the publication context may end on a concept and we consider the generalization of this concept. It may also end on a instance of concept and we consider all other instances of it. The path may also end on a value (from a data type property) inserted by the resource provider and we consider all the possible values.

5. Conclusion

In this paper we have presented an approach for indexing resources in a peer to peer network, where users are interested in their sharing. The resources are in every case semantically indexed via domain specific ontologies downloaded from the network. The semantic information strictly related to a resource represents a point of view of the user on the resource. The semantic index, which can be considered as a distributed RDF knowledge base is inserted in a Distributed Hash Table whose structure guaranties an efficient management of the resources. We consider that the ontologies required by the indexing can be provided by some expert users.

It is necessary to navigate the suitable ontologies in order to understand their different concepts and relations. This operation is time consuming but, nevertheless, it is the price to pay when using a semantic indexing and for profiting of its advantages regarding a keyword indexing. Ontologies allow some reasoning and we have shown how the structure of the index requires taking into account this characteristic for improving the queries. The proposed solution consists in foreseeing during the publication of a resource, different reasonable retrieval situations, and therefore different queries to which the resource should respond positively.

References

Baader F, Sattler U (2001) Tableau Algorithms for Description Logics. Studia Logica: An International Journal for Symbolic Logic.

David Beckett, Tim Berners-Lee (2008) Turtle - Terse RDF Triple Language. W3C Team Submission http://www.w3.org/TeamSubmission/turtle/.

Ghebghoub, O, Abel, M.-H, & Moulin, C (2008, July 1-5). *Learning Object Indexing Tool Based on a LOM Ontology.* Eighth IEEE International Conference on Advanced Learning Technologies, 2008. ICALT '08., Santander, Spain.

Ghebghoub, O, Abel, M-H, Moulin, C, & Leblanc, A (2009, June 12-14). *A LOM ontology put into practice.* Second International Conference on Web and Information Technologies, ICWIT 2009, Kerkennah Island Sfax, Tunisia.

Gruber T R (1993) Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer. Kluwer Academic Publishers.

Jing Mei, Parsia B (2004) Reasoning Paradigms for OW Ontologies. Department of Information Science, Freie Universitat Berlin, Techreport.

Kiryakov A, Ognyanov D, Manov D (2005) OWLIM - A Pragmatic Semantic Repository for OWL WISE Workshops. pp 182-192.

Moulin, C, Barthès, J-P, Bettahar, F, & Sbodio, M (2008, April 23-25). *Representation of Semantics in an E-Government Platform.* 6th Eastern European eGovernment Days, Prague, Czech Republic

Pan Z (2005) Benchmarking DL Reasoners Using Realistic Ontologies. Proceedings of the OWLED*05 Workshop on OWL: Experiences and Directions.

Parsia B, Sirin E (2004) Pellet: An OWL DL Reasoner. In 3rd International Semantic Web Conference (ISWC2004).

Salton G, Fox Edward A, Wu Harry (1982*)* *Extended Boolean information retrieval.* Technical Report, Cornell University.

Automatic acquisition of synonyms for French using parallel corpora

Lonneke van der Plas, Joerg Tiedemann, Jean-Luc Manguin

Abstract In this paper we describe an approach to the automatic extraction of synonyms for French that is easy to port across domains and across languages. The approach relies on automatic word alignments in parallel texts and uses distributional methods to compute the semantic similarity of words based on these word alignments. As a result the system outputs ranked lists of candidate synonyms for a given word. We compare the performance with a system that uses syntactic contexts to acquire synonyms automatically. Evaluations are done on a large-scale French synonym dictionary. We show that the alignment-based method outperforms the syntactic method by a large margin. In addition we show that the method can easily be ported to a different language and to a different domain.

1 Introduction

Support for semantics has been mentioned in the call for papers of the DART workshop 2010 and elsewhere as one of the goals of next generation information retrieval tools. Knowledge about synonymy is a type of lexico-semantic knowledge that can be useful, for example for search engines. Imagine a French student that searches for a job and types *cherche boulot* into Google. The student might be ignorant to the fact that the word *boulot* is a colloquial term for synonyms such as *travail, poste*. However, these terms (*travail, poste*) are used in the large majority of job announcements. Hence, simple

Lonneke van der Plas
University of Geneva, Switzerland, e-mail: lonneke.vanderplas@unige.ch
Jörg Tiedeman
Uppsala University, Sweden, e-mail: jorg.tiedemann@lingfil.uu.se
Jean-Luc Manguin
CNRS/University of Caen, France, e-mail: jean-luc.manguin@unicaen.fr

word matching will not retrieve the information needed by the student. Resources that group French synonyms could help this user in fulfilling his/her information need.

Synonym dictionaries are a common source of semantic information that could be able to deal with the problem described above. However, apart from semantics, personalisation and context awareness have been mentioned as goals to improve information retrieval tools. The drawback of dictionaries is that they are static and based on common knowledge, and therefore they are not personalised. Providing domain-dependent lexical information is a first step towards context-aware search engines. Search engines need to know when to relate a word like *bank* with the establishment for the custody of money (in the financial domain, for example) and when to relate it to the shore of a river. There are domain-specific dictionaries but the number of domains covered is limited.

Automatic methods for synonym acquisition are more flexible and therefore more easily adjustable to emerging needs. For example, recent work on the acquisition of synonyms using distributional models has shown that syntactic contexts can be applied to any large corpus of text that is analysed syntactically to acquire semantically related words [13, 17] and the method has been applied to corpora from different domains [21]. However, one of the prerequisites of this method is a large parsed corpus or at least a syntactic parser for the target language. For English there are many parsers available but for the majority of languages such tools do not exist. A personalised search tool would at least want to serve the user in his/her own language. Therefore, we need automatic methods that are easily ported to different languages and domains and that do not rely on language-specific pre-processing.

In this paper we will present a method that is particularly well-suited to be ported across different languages and across different domains. Moreover, the method outperforms the syntax-based approach described above for the task of synonym acquisition even when using smaller amounts of data.

Before we move to describing the methodology we need to explain the hypothesis that underlies our work. It is the distributional hypothesis. The hypothesis states that semantically related words are distributed similarly over contexts [12]. In other words, you can grasp the meaning of a word by looking at its contexts.

Context can be defined in many ways. Previous work has been mainly concerned with the syntactic contexts a word is found in. For example, the verbs that are in a subject relation with a particular noun form a part of its context. These contexts can be used to determine the semantic relatedness of words. For instance, words that occur in a object relation with the verb *to drink* have something in common: they are liquid.

Yet another context that is much less studied in vector-based approaches is the translational context. The translational context of a word is the set of translations it gets in other languages. For example, the translational context of *cat* is *kat* in Dutch and *chat* in French. How do we get from translational contexts to synonymy? The idea is that words that share a large number of translations are similar. For example both *autumn* and *fall* get the translation *herfst* in Dutch, *Herbst* in German, and *automne* in French. This indicates that *autumn* and *fall* are synonyms.

A straightforward place to start looking for translational context is in bilingual dictionaries. However, these are not always publicly available for all languages. More importantly, dictionaries are static and often incomplete resources, and they do not provide frequency information. We have chosen to automatically acquire word translations in multiple languages from text. Text in this case should be understood as multilingual parallel text. Automatic alignment gives us the translations of a word in multiple languages. Any multilingual parallel corpus can be used for this purpose. It is thus possible to focus on a special domain. Furthermore the automatic alignment provides us with frequency information for every translation pair, which can be handy in case words are ambiguous.

In this paper we compare a method that uses the translational context to a method that uses the syntactic context to determine the distributional similarity of words. In addition we will show how the method can be easily extended to different domains and different languages.

2 Alignment-based distributional similarity

In this section we explain the alignment-based approaches to distributional similarity. We will give some examples of translational context and we will explain how measures serve to determine the similarity of these contexts. We end this section with a discussion of related work.

2.1 Translational context

The translational context of a word is the set of translations it gets in other languages. For the acquisition of translations for French words we rely on automatic word alignment in parallel corpora.

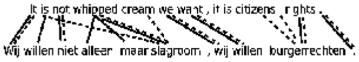

It is not whipped cream we want , it is citizens r ghts .
Wij willen niet allear maar slagroom , wij willen burgerrechten .

Fig. 1 Example of bidirectional word alignments of two parallel sentences

Figure 1 illustrates the automatic word alignment between a Dutch and an English phrase as a result of using the IBM alignment models [5] implemented in the open-source tool GIZA++ [19]. The alignment of two texts is bidirectional. The Dutch text is aligned to the English text and vice versa (dotted lines versus continuous lines). The alignment models produced are asymmetric. Several heuristics exist to combine directional word alignments. The intersection heuristics, for example, only accepts translation pairs that are found in both directions.

2.2 Measures for computing similarity

Translational co-occurrence vectors are used to find distributionally similar words. We give an example in Table 1. Every cell in the vector refers to a particular translational co-occurrence type. For example, *chat* ‘cat’ gets the translation *Katze* in German. The value of these cells indicate the number of times the co-occurrence type under consideration is found in the corpus.

Each co-occurrence type has a cell frequency. Likewise each head term has a row frequency. The row frequency of a certain head term is the sum of all its cell frequencies. In our example the row frequency for the term *chat* ‘cat’ is 64. Cut-offs for cell and row frequency can be applied to discard certain infrequent co-occurrence types or head terms respectively.

	Arbeit-DE	baan-NL	lavoro-IT	job-EN	cat-EN	Katze-DE
poste	17	26	8	13	0	0
boulot	6	12	7	10	0	0
chat	0	0	0	0	26	34

Table 1 Translational co-occurrence vector for *poste* ‘job’ *boulot* ’job’, and *chat* ’cat’ based on four languages

The more similar the vectors are, the more distributionally similar the head terms are. We need a way to compare the vectors for any two head terms to be able to express the similarity between them by means of a score. Various measures can be used to compute the distributional similarity between terms. We will explain in section 3 what measures we have chosen in the current experiments.

Furthermore, it has been shown that distributional methods benefit from using feature weights. For example in syntax-based approaches selectionally weak [25] or *light* verbs such as *hebben* ‘to have’ are given a lower weight than a verb such as *uitpersen* ‘squeeze’ that occurs less frequently. We have used the same weights for the translational context to counter balance the alignment errors that often occur with frequent words.

2.3 Related work

Multilingual parallel corpora have been used for tasks related to word sense disambiguation such as target word selection [10] and separation of senses [26, 11, 15].

Some researchers present methods for the automatic acquisition of paraphrases [3, 14, 28, 2, 6]. The first two of these have used a monolingual parallel corpus to identify paraphrases. The last three employ multilingual corpora from which the last two are also based on automatic word alignment as our approach is as well.

Improving the syntax-based approach for synonym identification using bilingual dictionaries and parallel corpora has been discussed in [18], [32], [22], and [23]. The last study is on French synonym acquisition.

3 Materials and methods

In the following subsections we describe the setup for our experiments.

3.1 Data collection

For the alignment method we need a parallel corpus of reasonable size with French either as source or as target language. Furthermore, we would like to experiment with various languages aligned to French. The freely available Europarl corpus [16] includes 11 languages in parallel, it is sentence aligned [31], and it is of reasonable size. Thus, for acquiring French synonyms we have 10 language pairs with French the source language: Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), Finnish (FI), Dutch (NL), Italian (IT), Portuguese (PT), and Swedish (SV). We applied a lemmatiser [27] to the French part of the language pairs in order to 1) reduce data sparseness, and 2) to facilitate our evaluation based on comparing our results to existing synonym databases.

Context vectors are populated with the links to words in other languages extracted from automatic word alignment. We applied GIZA++ and the intersection heuristics as explained in section 2.1. From the word-aligned corpora we extracted translational co-occurrence types, pairs of source and target words in a particular language with their alignment frequency attached. Each aligned target word is a feature in the (translational) context of the source word under consideration. We removed word type links that include non-alphabetic characters to focus our investigations on real words and we transformed all characters to lower case.

Note that we rely entirely on automatic processing of our data. Thus, the results from automatic tagging, lemmatisation and word alignment include errors. [2] show that when using manual alignment the percentage of correct paraphrases significantly rises from 48.9% to 74.9%.

3.2 Comparing vectors

To compare the vectors of the terms we need a similarity measure. We have chosen to describe the functions used in this paper using an extension of the notation used by [17], adapted by [8]. Co-occurrence data is described as tuples: ⟨word, language, word′⟩, for example, ⟨chat, EN, cat⟩.

Asterisks indicate a set of values ranging over all existing values of that component of the relation tuple. For example, $(w, *, *)$ denotes for a given word w all translational contexts it has been found in in any language. For the example of *chat* in, this would denote all values for all translational contexts the word is found in: *Katze_DE*:17, *chat_FR*:26 etc. Everything is defined in terms of co-occurrence data with non-zero frequencies. The set of attributes or features for a given corpus is defined as:

$$(w, *, *) \equiv \{(r, w')|\exists(w, r, w')\}$$

Each pair yields a frequency value, and the sequence of values is a vector indexed by $r{:}w'$ values, rather than natural numbers. A subscripted asterisk indicates that the variables are bound together:

$$\sum (w_m, *_r, *_{w'}) \times (w_n, *_r, *_{w'})$$

The above refers to a dot product of the vectors for term w_m and term w_n summing over all the $r{:}w'$ pairs that these two terms have in common. For example we could compare the vectors for *chat* and some other term by applying the dot product to all bound variables.

We explained in 2.2 that some attributes contain more information than other attributes. We want to account for that using a weighting function, that will modify the cell values. There is a placeholder for the weighting function:

$$\sum \mathit{weight}(w_m, *_r, *_{w'}) \times \mathit{weight}(w_n, *_r, *_{w'})$$

This is an abbreviation of:

$$\sum_{(r,w')\in(w_m,*,*)\cap(w_n,*,*)} \mathit{weight}(w_m, r, w') \times \mathit{weight}(w_n, r, w')$$

We have limited our experiments to using Dice†, a variant of Dice, and Pointwise mutual information (MI, [7]) since they performed best in a large-scale evaluation experiment reported in [9]. Dice† is defined as:

$$Dice\dagger = \frac{2 \sum min(weight(W1, *_r, *_{w'}), weight(W2, *_r, *_{w'}))}{\sum weight(W1, *_r, *_{w'}) + weight(W2, *_r, *_{w'})}$$

Note that Dice † gives the same ranking as the well-known Jaccard measure, i.e. there is a monotonic transformation between their scores. Dice † is easier to compute and therefore the preferred measure [9].
Pointwise mutual information (MI) measures the amount of information one variable contains about the other. MI is computed as follows:

$$MI = log \frac{P(w, r, w')}{P(w, *, *)P(*, r, w')}$$

Here, $P(w, r, w')$ is the probability of seeing *chat* aligned to *the* in a French-English parallel corpus, and $P(w, *, *)P(*, r, w')$ is the product of the probability of seeing *chat* aligned to any word in the corpus and the probability of seeing *the* aligned to any word in the corpus.

4 Evaluation

There are several evaluation methods available to assess lexico-semantic data. [8] distinguishes several. We decided to compare against a gold standard, because there is a large French synonym dictionary available. We evaluated our results on the *Dictionnaire Electronique des Synonymes* (DES, [24]), which is based on a compilation of seven French synonym dictionaries. It contains 49,149 nodes connected by 200,606 edges that connect synonymous words.

We compare our results to the syntax-based method for French by [4]. [4] present an explorative study of using distributional similarity to extract synonyms for French. They use two corpora: the 200 million-word corpus of newspaper text from *Le Monde*, and a 30 million-word corpus consisting of 515 twentieth century novels. Several syntactic relations are extracted.

The test set was chosen by looking at the pairs of nearest neighbours resulting from the syntax-based method that receive a score not lower than 0.16. This resulted in a list of approximately 1000 nouns. Of this list 950 can be found in the data of the alignment-based method. This list of 950 word constitutes the test set.

5 Results and Discussion

The similarity scores calculated by the system for a pair of nearest neighbours is used as a threshold. Precision and recall are calculated as well as the coverage of the system at varying thresholds. The results can be seen in Figure 2 and Figure 3.

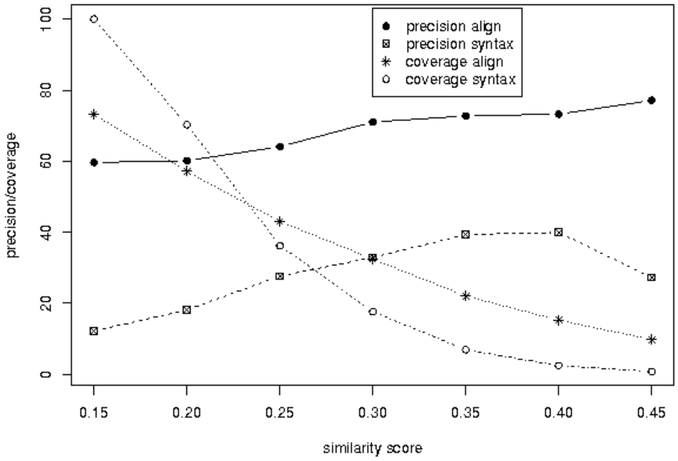


Fig. 2 Number of co-occurrence types when augmenting the cell frequency cutoff

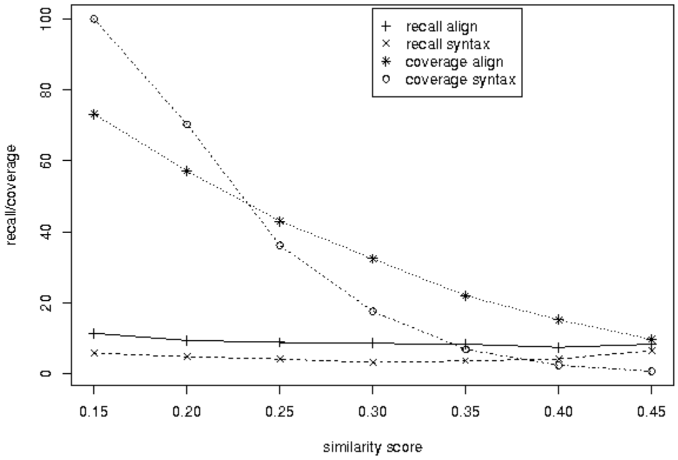


Fig. 3 Number of co-occurrence types when augmenting the cell frequency cutoff

Coverage of both systems decreases when the threshold for the similarity score is augmented. That is expected since not many words have nearest neighbours with a high similarity score. The alignment-based method never reaches 100% coverage. However, it should be noted that the test set was chosen in a way that favours the syntax-based method. The test set was chosen from the pairs of nearest neighbours resulting from the syntax-based method above the threshold 0.16. Thus, the coverage of the syntax-based method is 100% at 0.16. The coverage of the alignment-based method is approximately 70% for that threshold. However, the coverage of the syntax-based method decreases more rapidly as the thresholds are raised. At threshold 0.45 the coverage of the syntax-based method is close to zero.

If we compare the precision of the nearest neighbours for both systems at the same level of coverage (50%) we see that the syntax-based method has a precision score of 25%, whereas the alignment-based method produces nearest neighbours with a precision of 60% to 65%. The precision of the alignment-based method ranges between a little under 60% to a little under 80% at threshold 0.45. The precision of the syntax-based method ranges between 10% at threshold 0.16 and a little under 40% for threshold 0.4. It is striking that the precision drops at the end of the line, when the threshold is set to 0.45. The nearest neighbours with the highest scores are not the best. However, it should be noted that due to limited coverage (close to 0) the numbers at this threshold are unreliable.

With respect to recall, it can be concluded that there is a smaller difference between the two methods and the scores are less satisfactory in general. It should be noted that the dictionaries often include synonyms from colloquial language use. We do not expect to find these synonyms in the Europarl corpus. The advantage of corpus-based methods is that we are free to select a corpus that is most suitable for the task at hand.

A closer inspection of the nearest neighbours resulting from the alignment-based method, shows that many of the candidate synonyms judged incorrect are in fact valuable additions, such as *sinistre* ‘disaster’ for *accident* ‘accident’.

Many errors stem from the fact that the alignment-based method does not take multiword units into account. For the French data this typically results in many related adjectives and adverbs being selected as nearest neighbours. For example, *majoritaire* ‘majority (adj)’ is returned as a synonym for *majorité* ‘majority (noun)’, stemming from the multiword unit *parti majoritaire*. Also *majoritairement* and *largement* are among the nearest neighbours. Words that would be translated in English as *for the most part*. These translations that are composed of multiple words cause problems for the alignment method and hence for the synonyms extracted.

6 Porting to a different language

We explained in the introduction that the method easily ports to different languages. As an example we include results of applying the alignment-based method to Dutch taken from [21]. Instead of collecting translations for all French words, we collected translations for all Dutch words in the same parallel corpus used to acquire the French synonyms.

We post-processed the alignment results in various ways for Dutch. We applied a simple lemmatiser to the Dutch part of the bilingual translational co-occurrence types. For this we used two resources: CELEX, a linguistically annotated dictionary of English, Dutch, and German [1], and the Dutch snowball stemmer implementing a suffix-stripping algorithm based on the Porter stemmer. We removed word type links that include non-alphabetic characters and we transformed all characters to lower case.

Evaluations for Dutch were done on a large test set of 3000 nouns selected from Dutch EuroWordNet. We have split up the test set in high-frequency (HF), middle-frequency (MF) and low-frequency (LF) words. This was done to able to study the effect of frequency on the performance of the system.

Because the test set is comprised of nouns only we applied an extra step in pre-processing. We restricted our study to Dutch nouns. Hence, we extracted translational co-occurrence types for all words tagged as nouns in CELEX. We also included words that are not found in CELEX[1].

6.1 Results on Dutch synonym extraction

Also for Dutch we compare the results with a syntax-based method that is described in [21]. There syntactic relations are extracted from a 500 million word corpus (TwNC, [20]), a corpus that is much larger than that used for the alignment-based method.

Method	HF		MF		LF	
	k=1	k=5	k=1	k=5	k=1	k=5
Alignment-based	31.71	19.16	29.26	16.20	28.00	16.22
Syntax-based	21.31	10.55	22.97	10.11	19.21	11.63

Table 2 Percentage of synonyms over the k candidates for the alignment-based and syntax-based method for the three frequency bands

These results on the Dutch language show a similar pattern to the results on French. In spite of data sparseness, it is clear from Table 2 that the

[1] Discarding these words would result in losing too much information. We assumed that many of them will be productive noun constructions.

alignment-based method is better at finding synonyms than the syntax-based method. In Table 2 we see the percentage of synonyms found among the k candidate synonyms retrieved by the system.

7 Porting to a different domain

As a first step in the direction of context-awareness in search tools, we acquire domain-dependent synonyms from a corpus that is very different in nature from the Europarl corpus: a multilingual corpus of subtitles [30, 29]. The complete corpus contains about 21 million aligned sentence fragments in 29 languages. We used all language pairs that include French, 23 language pairs in total.

The domain is different from the domain of the Europarl Corpus. There is a world of difference between the working day of a member of the European Parliament and the adventures of Nemo. Above all, movie subtitles consist mainly of transcribed speech. In principle this is the same for the Europarl Corpus. However these proceedings are edited and far less spontaneous than the speech data from the movies.

7.1 Some examples of synonyms from different domains

The most interesting question when switching to a new corpus is to see whether this change also leads to a new domain of synonyms. We will give some examples below for which the domain switch is easily spotted.

Testword	Corpus			
ami	Europarl	amitié	camarade	allié
'friend'		'friendship'	'comrad'	'ally'
	Subtitles	copain	pote	amie
		'friend'	'buddy'	'girlfriend'
fille	Europarl	fillette	enfant	filial
'girl'		'small girl'	'child'	'relative to a daughter'
	Subtitles	fillette	nana	fiiie
		'small girl'	'babe'	(mistake in optical character recognition)
malade	Europarl	patient	maladie	souffrant
'ill'		'patient'	'illness'	'suffering'
	Subtitles	souffrant	fou	dingue
		'suffering'	'crazy'	'crazy'

Table 3 Examples of nearest neighbours at the top-3 ranks for two corpora

In future work we plan to run a quantitative evaluation on these synonyms that were acquired from a corpus of subtitles. Based on results in previous work on the Dutch language [21] we expect that the subtitle corpus will not retrieve as many synonyms that can be found in the gold standard. There are three reasons for this. The first reason is that the corpus is smaller. The second reason is that the subtitle corpus is more noisy than the Europarl corpus, for example due to mistakes in optical character recognition. The third reason, and this might hold only for the Dutch study, is that the synonyms retrieved are often very colloquial and therefore possibly harder to find in gold standards. This might be different for the French evaluations, because the DES often includes synonyms from colloquial language use.

8 Conclusions

In this article we have shown that for the task of automatic synonym acquisition the alignment-based method outperforms the traditional syntax-based method by a very large margin. The precision is more than twice as high for the alignment-based method and it manages to find valuable additions not present in the dictionary. In addition we showed that the method can be easily ported across languages and domains. We showed encouraging results on acquisition of Dutch synonyms and we compared the synonyms retrieved from proceedings of the European Parlement with synonyms retrieved from a parallel corpus of subtitles.

References

1. Baayen, R., Piepenbrock, R., van Rijn, H.: The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia (1993)
2. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL) (2005)
3. Barzilay, R., McKeown, K.: Extracting paraphrases from a parallel corpus. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 50–57 (2001). URL citeseer.ist.psu.edu/barzilay01extracting.html
4. Bourigault, D., Galy, E.: Analyse distributionnelle de corpus de langue générale et synonymie. In: Lorient, Actes des Journées de la Linguistique de Corpus (JLC) (2005)
5. Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2), 263–296 (1993)
6. Callison-Burch, C.: Syntactic constraints on paraphrases extracted from parallel corpora. In: Proceedings of EMNLP (2008)
7. Church, K., Hanks, P.: Word association norms, mutual information and lexicography. Proceedings of the Annual Conference of the Association of Computational Linguistics (ACL) (1989)

8. Curran, J.: From distributional to semantic similarity. Ph.D. thesis, University of Edinburgh (2003)
9. Curran, J., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 222–229 (2002)
10. Dagan, I., Itai, A., Schwall, U.: Two languages are more informative than one. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (1991)
11. Dyvik, H.: Translations as semantic mirrors. In: Proceedings of Workshop Multilinguality in the Lexicon II (ECAI) (1998)
12. Harris, Z.: Mathematical structures of language. Wiley (1968)
13. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) (1990)
14. Ibrahim, A., Katz, B., Lin, J.: Extracting structural paraphrases from aligned monolingual corpora. In: Proceedings of the second international workshop on Paraphrasing (IWP), pp. 57–64 (2003)
15. Ide, N., Erjavec, T., Tufis, D.: Sense discrimination with parallel corpora. In: Proceedings of the ACL Workshop on Sense Disambiguation: Recent Successes and Future Directions. (2002)
16. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the MT Summit, pp. 79–86. Phuket, Thailand (2005)
17. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of COLING/ACL (1998)
18. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying synonyms among distributionally similar words. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2003)
19. Och, F.: GIZA++: Training of statistical translation models. Available from `http://www.isi.edu/~och/GIZA++.html` (2003)
20. Ordelman, R.: Twente nieuws corpus (TwNC). Parlevink Language Techonology Group. University of Twente. (2002)
21. van der Plas: Automatic lexico-semantic acquisition for question answering. Groningen dissertations in linguistics (2008)
22. van der Plas, L., Tiedemann, J.: Finding synonyms using automatic word alignment and measures of distributional similarity. In: Proceedings of COLING/ACL (2006)
23. van der Plas, L., Tiedemann, J., Manguin, J.: Extraction de synonymes à partir d'un corpus multilingue aligné. In: Actes des 5èmes Journées de Linguistique de Corpus à Lorient (2008)
24. Ploux, S., Manguin, J.: Dictionnaire électronique des synonymes français (1998, released 2007)
25. Resnik, P.: Selection and information (1993). Unpublished doctoral thesis, University of Pennsylvania
26. Resnik, P., Yarowsky, D.: A perspective on word sense disambiguation methods and their evaluation. In: Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, what, and how? (1997)
27. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, pp. 44–49. Manchester, UK (1994). Http://www.ims.uni-stuttgart.de/~schmid/
28. Shimota, M., Sumita, E.: Automatic paraphrasing based on parallel corpus for normalization. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) (2002)
29. Tiedemann, J.: Building a multilingual parallel subtitle corpus. In: Proceedings of the Conference on Computational Linguistics in the Netherlands (CLIN) (2007)
30. Tiedemann, J.: Improved sentence alignment for building a parallel subtitle corpus. In: Proceedings of the Conference on Computational Linguistics in the Netherlands (CLIN) (2007)

31. Tiedemann, J., Nygaard, L.: The OPUS corpus - parallel & free. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) (2004)
32. Wu, H., Zhou, M.: Optimizing synonym extraction using monolingual and bilingual resources. In: Proceedings of the International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP) (2003)