Exploiting Lexical Resources for Therapeutic Purposes: the case of WordNet and STaRS.sys

Gianluca E. Lebani[†‡]
[†]Center for Mind/Brain Sciences
University of Trento
gianluca.lebani@unitn.it

Emanuele Pianta[‡]
[‡]HLT Group
Fondazione Bruno Kessler
pianta@fbk.eu

Abstract

In this paper, we present an on-going project aiming at extending the WordNet lexical database by encoding common sense featural knowledge elicited from language speakers. Such extension of WordNet is required in the framework of the STaRS.sys project, which has the goal of building tools for supporting the speech therapist during the preparation of exercises to be submitted to aphasic patients for rehabilitation purposes. We review some preliminary results and illustrate what extensions of the existing WordNet model are needed to accommodate for the encoding of commonsense (featural) knowledge.

1 Introduction

Electronic lexical resources such as WordNet and FrameNet are used for a great variety of natural processing tasks, ranging from query expansion, to word sense disambiguation, text classification, or textual entailment. Some of these resources are also used by human users as on-line dictionaries; see the Princeton WordNet[1] and the MultiWordNet[2] on-line sites. In this paper we describe a novel attempt to exploit the information contained in wordnets to build a tool designed to support the therapy of language disorders. In doing so, we will tackle also an interesting theoretical issue. Is the WordNet conceptual model apt to represent the common sense knowledge associated to concepts, which is partly lost in case of language disorders (aphasia) due to a brain damage? Note that, in cognitively oriented studies of the lexicon such knowledge is often represented in the form of featural descriptions elicited from speakers, such as <*a cat*> *is lazy*[3], <*camels*> *are found in deserts*, <*planes*> *fly* etc.

Anomia is the most pervasive and persistent of aphasia symptoms. It has been described as “a difficulty in finding high information words, both in fluent discourse and when called upon to identify an object of action by name” (Goodglass and Wingfield, 1997:3). The naming difficulties experienced by anomic patients can vary substantially, so that different “anomias” can be characterized as arising from either a mainly lexical or mainly semantic breakdown. Depending on the kind of anomia, therapeutic approaches can vary, so as to employ the more appropriate tasks and stimuli.

Computers can support the rehabilitation of language disorders in many ways: from assisting the administrative management to enhancing common assessment methods, from helping the clinician during the therapeutic session to alleviating the communicative difficulties of a patient by exploiting his unimpaired abilities (Petheram, 2004).

In these pages we introduce STaRS.sys (Semantic Task Rehabilitation Support system), a Computer Assisted Therapy (CAT) tool designed to support the therapist in the preparation of semantic exercises such as odd-one-out, yes/no attribute question answering, property generation and so forth. All these exercises are based on the kinds of information that are carried by featural

[1] http://wordnet.princeton.edu/
[2] http://multiwordnet.fbk.eu/

[3] Concepts and features will be printed in *italics courier new* font. When reporting a concept-feature pair, the concept will be further enclosed by <*angled brackets*>. Feature types and concept categories will be reported in *italics times new roman.*

descriptions. Such a scenario motivates the need for a lexical semantic resource which is richer and somehow more cognitively-oriented than the existing ones. We will argue that such needs can be satisfied by enhancing the WordNet model (WN: Fellbaum, 1998 ed) as implemented in the Italian MultiWordNet (MWN: Pianta et al, 2002) lexicon. Our project is developed in collaboration with the CIMeC's Center for Neuropsychological Rehabilitation (CeRiN), and focuses on Italian. We leave to the future the evaluation of whether and how our model can be expanded to other languages.

These pages are organized as follows: Sec. 2 shows the possibilities offered by the exploitation of STaRS.sys in a therapeutic context, and the lexical semantics requirements that such use poses. In Sec. 3 and 4 we illustrate specific issues related to the encoding of featural knowledge into the MWN model.

2 STaRS.sys in a therapeutic context

In this section we will illustrate the semantic requirements that the therapeutic use of STaRS.sys poses, and how we foresee the tool will be used in practical therapeutic scenarios.

2.1 Semantic requirements

An essential requirement of the STaRS.sys tool is the capability of managing the major variables that influence the performance of anomic patients in semantic therapeutic tasks (Raymer and Gonzalez-Rothi, 2002; Cree and McRae, 2003). Accordingly, we identified a minimum of five types of information which should be available for every lexical concept:

Conceptual Taxonomy. A fully-specified conceptual taxonomy is an essential requirement for our tool, in the light of the existence of patients affected by language disorders specific to certain semantic categories, such as *tools*, or *living beings* (Capitani et al, 2003).

Featural Descriptions. Featural descriptions are assumed to play a central role in the human semantic memory (Murphy, 2002) and will be represented here as *<concept> feature* couples, e.g. *<dog> has a tail*.

This information can be exploited for selecting sets of concepts which are relevant in a certain therapeutic context, e.g. concepts sharing a feature value ("red objects") or those for which a type of feature is particularly relevant (e.g. "animals with a peculiar fur").

Feature Types Classification. A grouping of FDs into feature types is needed for selectively working on feature types of interest, or for the estimation of semantic measures such as feature distinctiveness, semantic relevance, concept similarity and feature correlation (Cree and McRae, 2003; Sartori and Lombardi, 2004; Vinson et al, 2003). As we will see in the following sections, feature types can be mapped onto WordNet-like relations.

Prototypicality. A concept can be more or less representative of its category. Choosing and working on concepts with different levels of prototypicality can be informative, for both therapeutic and diagnostic purposes.

Word Frequency. Patients' performance can be affected by word frequency. Thereby, a critical skill for our tool is the ability to discriminate between words used with different frequencies.

2.2 Use Case Scenarios

By exploiting a lexical infrastructure encoding such semantic information, STaRS.sys can be used by a therapist for:

- retrieving concepts;
- retrieving information associated to concepts;
- comparing concepts.

These three functionalities can be illustrated by the preparation of three different tasks for a patient affected by, e.g., a semantic deficit selectively affecting animal concepts. Such a kind of patient would show comprehension and production difficulties restricted to concepts belonging to the *animal* category (Capitani et al, 2003). Plausibly, furthermore, his production problems would manifest both as naming failure in controlled conditions (i.e. in tests like the ones reported below) and as a difficulty/inability to retrieve the intended word in spontaneous speech (Semenza, 1999).

In the first scenario, the therapist looks for concepts that match given specifications in order to prepare a feature generation task. As an example, she submits to STaRS.sys a request for concepts of frequent use, referring to animals, associated to highly distinctive color features and having a high mean feature distinctiveness. The system returns concepts such as *zebra*, *tiger* and `cow`. Finally the patient is asked to generate phrasal descriptions for these concepts.

In a second scenario, STaRS.sys is used to retrieve FDs for a given set of concepts. Right and wrong concept-feature couples are created to build a questionnaire, in which the patient is required to distinguish the right from the wrong pairs. For instance, the therapist submits to STaRS.sys a query for features of the concept `leopard` that are highly relevant and either perceptual or taxonomical, and obtains features such as `is yellow with black spots` and `is a cat`.

Finally, in the third scenario the therapist uses STaRS.sys to find concepts for an odd-one-out task. That is, she looks for triples composed of two similar concepts plus an incoherent one that has to be found by the patient. As an example, starting from the concept `lion`, she looks for animals that typically live in a similar/different natural habitat, and obtains similar concepts such as `leopard` and `cheetah`, and a dissimilar concept such as `wolf`.

3 WN as semantic lexical resource for STaRS.sys

The STaRS.sys application scenario motivates the need for a lexical semantic resources that:

R1: is cognitively motivated;

R2: is based on a fully-specified is-a hierarchy;

R3: is intuitive enough to be used by a therapist;

R4: allows for the encoding of featural properties and their association to concepts;

While designing the STaRS.sys tool, we made the hypothesis that a semantic lexical resource built according to the WN model could meet most of the above requirements.

In the WN model every concept is represented as a synset (set of synonyms) such as {hand, manus, hook, mauler, mitt, paw}. Such semantic units are organized in a network interconnected trough several relations. Examples of semantic relations include the *is-a* relation, e.g.{left_hand, left} *is-a* {hand, …}, and the *meronymy* relation, e.g. {hand, …} *has-part* {finger}.

At a first glance, WN seems to easily meet three of the above criteria. First, WN was initially conceived as a model of the human lexical memory. Second, WordNet implement extensive and systematic noun hierarchies. More specifically, a preliminary analysis of the Italian MWN nominal hierarchy has shown that the semantic categories which are relevant for rehabilitation purposes can be easily mapped onto MWN top level nodes (*tools*, *animals*, *people*). Third, WN is based on a conceptual model which is relatively simple and near to language use (as opposed to more sophisticated logics-based models). We expect that this feature will facilitate the use of STaRS.sys by therapists, which may not have all the formal logics awareness that is needed to use formal ontologies. Furthermore, MWN is manually developed trough an on-line Web application. We expect that such application can be used by therapists using STaRS.sys for the shared and community-based development/maintenance of the lexical resource they need.

A final motivation in favor of the choice of MWN is the fact that this Italian resource is strictly aligned at the semantic level to English and other European languages (e.g. Spanish, Portuguese, Romania, Hebrew). Thus, we can envisage that at least part of the semantic information which is encoded for Italian can be ported to the aligned languages and used for similar purposes.

4 Mapping featural descriptions into MWN

Our hypothesis about the usefulness of the WN model for the needs of STaRS.sys can be fully confirmed only if we find a way to encode in such a model all or most of the knowledge which is contained in feature descriptions elicited from Italian speakers (R4 in previous section). In more general terms we need to answer the following questions. Does MWN already contain all the information that is needed by the STaRS.sys requirements? If we need to extend the existing MWN, can we simply add new synsets and instances of existing relations, or do we need to add new relation types? Is the conceptual model of MWN or of any other WN variant powerful enough to encode all the information contained in feature descriptions?

A first simple approach to representing feature descriptions in MWN is associating feature descriptions to synset glosses. As a consequence, a MWN gloss, which is currently composed of a definition and a list of usage examples, all crafted by lexicographers, would contain also a list of feature descriptions, elicited from language speakers.

This approach may be useful for some of the foreseen usages of STaRS.sys (e.g. retrieving feature descriptions from concepts), and can also be interesting for a generic use of MWN. However, to fully exploit the knowledge contained in FDs (e.g. for calculating concept similarity) it is necessary to encode that knowledge in a more explicit way; that is we need to map each FD in a *wordnet-like relation* between a *source* and a *target* concept. For instance, a pair such as *<`cup`>* *`is used for drinking`* can be represented as a *is_used_for* relation holding between the source concept {cup} and the target concept {drink}.

Encoding the source concept is relatively easy given that it is usually expressed as an isolated word that is used as stimulus for feature elicitation from subjects, e.g. “scimmia” (“monkey”). The only problematic aspect in this step may be the choice of the right sense which was meant when the word has been proposed to subjects. In some cases this may be not trivial, even if, in principle, stimulus words are supposed to be chosen so as to avoid ambiguities; see for instance the word “cipolla” (“onion”), which in MWN is ambiguous between the vegetable and food sense.

More complex is the encoding of the feature itself which is a free and possibly complex linguistic description (e.g. `likes eating bananas`). To fulfill our goal, we need to map such description in a wordnet-like relation and a target concept. Such goal can be accomplished in two steps.

4.1 Mapping feature types into MWN relations

Given the semantic requirements illustrated in Sec. 2.1, one the first steps in the development of the STaRS.sys tool has been the design of a classification of FDs in feature types; see Lebani and Pianta (2010). In a second moment, we realized that assigning a FD to a feature type is equivalent to assigning it to a wordnet relation, given that it is possible to create one-to-one mappings between features types and relations.

The adopted feature type classification has been designed so as to be (1) reasonably intuitive, (2) robust and (3) cognitively plausible. The cognitive plausibility requirement has been fulfilled by moving from an analysis of similar proposals put forwards in the experimental literature, or exploited in the therapeutic practice. As for the former, we considered research fields as distant as lexicography, theoretical linguistics and cognitive psychology. Examples of compatible proposals currently exploited in the therapeutic practice are the question type of Laiacona et al’s (1993) semantic questionnaire, a type classification adopted by the therapists of the CIMeC’s CeRiN (personal communication) and the Semantic Feature Analysis paradigm (Boyle and Coelho, 1995).

The resulting classification only considers concrete objects and is composed of 25 feature types. All of them (except the *is associated with* relations) belong to one of the following six relations) belong to one of the following six major classes: taxonomic properties, part-of- relations,

Feature Type	Example
has Portion	*<bread> cut into slices*
has Geographical Part	*<Africa> Egitto*
has Size	*<elephant> is big*
has Shape	*<clock> is round*
has Texture	*<eel> is slimy / <biscuit> is crunchy*
has Taste	*<lemon> is bitter*
has Smell	*<rose water> smells of rose*
has Sound	*<lighting> produces a thunder*
has Colour	*<lemon> is yellow*
is Used for	*<cup> is used for drinking*
is Used by	*<cleaver> is used by butchers*
is Used with	*<violin> is played with a bow*
Situation Located	*<jacket> used in occasions*
Space Located	*<camel> in the desert*
Time Located	*<pajamas> used at night*
has Origin	*<milk> comes from cows*
is Involved in	*<bird> eats seeds - is hunted*
has Attribute	*<subway> is fast*
has Affective Property	*<horror movie> is scary*
is Associated with	*<dog> man*

Table 1: STaRS.sys types not having a parallel wordnet semantic relation

perceptual properties, usage properties, locational properties and associated events and attributes.

A first version of this classification has been evaluated by asking 5 naïve Italian speakers to assign the appropriate type label to 300 concept[4]-feature pairs from a non-normalized version of the Kremer et al's (2008) norms. The inter-coder agreement between subjects (Fleiss' Multi-π = 0,73) validated the skeleton of our classification, at the same time suggesting some minor changes that have been applied to the classification proposed here. An evaluation of the improved classification involving therapists has been planned for the (very near) future.

Note that in order to map all of the feature types into wordnet relations we had to create a number of new relations which are not available in existing wordnets. The list of existing MWN relations used to encode STaRS.sys feature types includes five items: *hypernym, has_co-ordinate, has_part, has_member, has_substance*. The following table contains the list of the 20 additional relations, along with examples.

4.2 Encoding target concepts in MWN

A second step needed in order to fully represent the semantics of feature descriptions in MWN is the encoding of target concepts.

Target concepts can be expressed by a noun (e.g. *has a <neck>*), an adjective (e.g. *is <big>*) or a verb or a verbal construction (e.g. *is used for <drinking>*, *is used to <cut bread>*). In principle this is not problematic as WN encodes all these lexical categories.

What is problematic instead is the possible complexity of target concepts. Whereas in WN synsets are bound to contain only lexical units (with the few exceptions of the so called artificial nodes), the target of a featural description can be a free combination of words, for instance a noun modified by an adjective (e.g. *has a <long neck>*), an adjective modified by an adverb (e.g. *is <very big>*) or a verb with an argument (e.g. *is used to <cut bread>*). For giving an idea of the phenomenon, consider that 27,6% of the features that composes the experimental sample in Lebani and Pianta (2010) contain target concepts expressed by free combination of words

The solution we adopted to solve this problem relies on the notion of *phraset* proposed by Bentivogli and Pianta (2003; 2004), that is a data structure used for encoding "sets of synonymous free combination of words (as opposed to lexical units) which are recurrently used to express a concept". In the original proposal, the authors introduced such a data structure to cope with lexical gaps in multingual resources or to encode alternative (linguistically complex) ways of expressing an existing concept. Phrasets can be associated to existing synsets to represent alternative (non lexical) ways of expressing lexicalized concepts, e.g. the Italian translations of "dishcloth":

Synset: {canovaccio, strofinaccio}
Phraset: {strofinaccio_per_i_piatti,
straccio_per_i_piatti}

where "strofinaccio per i piatti" and "straccio per i piatti" and are free combinations of words. In alternative, they can be used to represent lexical gaps, such as the Italian translation equivalent of "breadknife":

Synset: {GAP}
Phraset: {coltello_da_pane,
coltello_per_il_pane}

Phrasets can be annotated by exploiting the *composes/composed-of* lexical relation linking phraset with the synsets corresponding to the concepts that compose it. For instance the expression in the above phraset is linked by a *hypernym* and by a *composed-of* relation with the synset {coltello} (*knife*) and {pane} (*bread*). As far as FDs are concerned, the use of phrasets is compatible with the received view about the compositional nature of the human conceptual knowledge (Murphy, 2002).

Figure 1 shows how phrasets allow for representing the complex FD *<breadknife> is used to cut bread* in the MWN model.

5 Conclusion and future directions

This paper presents the preliminary results of a research aiming at exploiting and extending the WordNet conceptual model as an essential component of a tool for supporting the rehabilitation of patients with language disorders. A crucial

[4] In details, the subjects were submitted with concrete concepts belonging to one of the following categories: *mammals, birds, fruits, vegetables, body parts, clothing, manipulable tools, vehicles, furniture* and *buildings*.

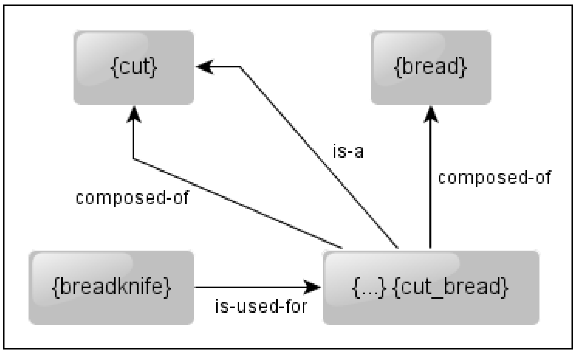

Figure 1: Representation of the concept-feature pair <*breadknife*> *is used to cut bread*

aspect for the use of wordnet-like resources in such a context is the possibility of representing lexical knowledge represented in the form of feature descriptions elicited from language speakers. Our work has illustrated the steps which are needed to encode feature descriptions in the WN model. To this purpose we introduced twenty new wordnet relations, and relied on phrasets for representing complex (non-lexicalized) concepts.

The study presented in these pages is a necessary theoretical step for the development of our tool. A practical evaluation of its feasibility is planned for the very near future, together with other (equally important but less relevant in this context) issues concerning both the population of our semantic knowledge base and the overall design of STaRS.sys.

Acknowledgements. We are grateful to Rita Capasso and Alessia Monti for the useful discussion of the application scenario sketched in these pages.

References

Luisa Bentivogli and Emanuele Pianta. 2003. Beyond Lexical Units: Enriching WordNets with Phrasets. *Proceedings of EACL 2003*: 67-70.

Luisa Bentivogli and Emanuele Pianta. 2004. Extending WordNet with Syntagmatic Information. *Proceedings of the 2nd International WordNet Conference*: 47-53.

Mary Boyle and Carl A. Coelho. 1995. Application of semantic feature analysis as a treatment for aphasia dysnomia. *American Journal of Speech-Language Pathology*, 4: 94-98.

Erminio Capitani, Marcella Laiacona, Brad Z. Mahon and Alfonso Caramazza. 2003. What are the Facts of Semantic Category-Specific Deficits? A Critical Review of the Clinical Evidence. *Cognitive Neuropsychology*, 20(3): 213-261.

George S. Cree and Ken McRae. 2003. Analyzing the Factors Underlying the Structure and Computation of the Meaning of Chipmunk, Cherry, Chisel, Cheese, and Cello (and Many Other Such Concrete Nouns). *Journal of Experimental Psychology: General*, 132 (2): 163-201.

Christiane Fellbaum. 1998 ed. *WordNet: an electronic lexical database*. The MIT Press.

Harold Goodglass and Arthur Wingfield. 1997. *Anomia: Neuroanatomical & Cognitive Correlates*. Academic Press.

Gerhard Kremer, Andrea Abel and Marco Baroni. 2008. Cognitively salient relations for multilingual lexicography. *Proceedings of COLING-CogALex Workshop 2008*: 94-101.

Marcella Laiacona, Riccardo Barbarotto, Cristina Trivelli and Erminio Capitani. 1993. Dissociazioni Semantiche Intercategoriali. *Archivio di Psicologia, Neurologia e Psichiatria*, 54: 209-248.

Gianluca E. Lebani and Emanuele Pianta. 2010. A Feature Type Classification for Therapeutic Purposes: a preliminary evaluation with non expert speakers. *Proceedings of ACL-LAW IV Workshop*.

Gregory L. Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.

Brian Petheram. 2004, ed. Special Issue on Computers and Aphasia. *Aphasiology*, 18 (3): 187-282.

Emanuele Pianta, Luisa Bentivogli and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. *Proceedings of the 1st International Conference on Global WordNet*.

Anastasia Raymer and Leslie Gonzalez-Rothi. 2002. Clinical Diagnosis and Treatment of Naming Disorders. In A.E. Hillis (ed) *The Handbook of Adult Language Disorders*. Psychology Press: 163-182.

Giuseppe Sartori and Luigi Lombardi. 2004. Semantic Relevance and Semantic Disorders. *Journal of Cognitive Neuroscience*, 16 (3): 439-452.

Carlo Semenza. 1999. Lexical-semantic disorders in aphasia. In G. Denes and L. Pizzamiglio (eds.). *Handbook of Clinical and Experimental Neuropsychology*. Psychology Press, Hove: 215-244.

David P. Vinson, Gabriella Vigliocco, Stefano Cappa and Simona Siri. 2003. The Breakdown of Semantic Knowledge: Insights from a Statistical Model of Meaning Representation. *Brain and Language*, 86: 347-365.