The Role of MultiWord Terminology in Knowledge Management

James Dowdall*, Will Lowe†, Jeremy Ellman‡, Fabio Rinaldi*, Michael Hess*

*Institute of Computational Linguistics, University of Zürich, Switzerland
{dowdall, hess, rinaldi}@cl.unizh.ch

†WordMap Ltd., Bath, UK
will.lowe@wordmap.com

‡School of Informatics, Northumbria University, UK
jeremy.ellman@northumbria.ac.uk

Abstract
One of the major obstacles for knowledge management remains MultiWord Terminology (MWT). This paper explores the difficulties that arise and describes real world solutions implemented as part of the Parmenides project. Parmenides is being built as an integrated knowledge management package that combines information, MWT and ontology extraction methods in a semi-automated framework. The focus of this paper is on eliciting ontological fragments based on dedicated MWT processing.

1. Introduction

In rapidly developing areas such as biotechnology, market competitors keep a close eye on novel market developments. Externally, newswire feeds pump facts, figures and opinions about mergers, launches and trends. Internally, R&D reports on experimental results, product development and consumer trends. However, before influencing the decision making process, this information needs to be integrated into a centralized knowledge base. With employees spread across languages and boarders, categorizing text often becomes subjective and culturally biased. (Semi) Automating this process produces not only quantitative gains but also qualitative improvements by enforcing heterogeneous, systematic categorization, defined against an ontology.

The state of the art allows for near perfect automatic categorization of Named Entities (NEs) such as people, organizations, dates and other MUC familiars. However, centralizing information gathering and analysis requires being as familiar with technical reports as with newswire. The pervasive use of MultiWord Terminology (MWT) that characterizes these reports, combined with an almost complete lack of NEs, necessitates additional computational effort to turn a domain’s ‘jargon’ into an indispensable knowledge source. This effort must tackle the two problems of MWT extraction and organization - determining which MWTs appear in a document and recognizing any ontological structure between them. This paper describes the solutions to these problems adopted within the Parmenides project1. The TermFinder system performs MWT extraction and identifies ontological links between MWTs.

2. MWT Extraction

MWT extraction is a developing field (Castellvi et al., 2001) and there exist many methods, but rather little theory to guide our choice among them (Kageura, 2002). In fact most successful methods, including the TermFinder, use a combination of techniques (hybrid approach). Until more theory is developed, current applications will continue to incorporate term extraction and organization as a semi-automatic process, making essential use of human judgments to validate the results.

Existing methods can be divided roughly according to the specificity of the knowledge sources used to find candidate MWTs and ontological links. The least domain-specific methods for term extraction use weighted document frequency counts to measure the probable importance of terms in the domain e.g. C/NC techniques (Frantzi and Ananiadou, 1996). More specific methods of term extraction use parse tree or part of speech information inferred from the document to single out subtree or POS patterns that imply the existence of a term e.g. the sequence “JJ NP” is a potential term as in “potential/JJ term/NP”.

3. Identifying Ontological Structure

Domain-general methods for ontology construction include clustering vectors of word co-occurrences to define similarity over terms; terms similar according to this metric can be treated as synsets. Linguistic information can also be utilized in extracting an ontology structure e.g. the pattern “NP1 such as NP2” may indicate an ISA relation between the entities named by the heads of NP1 and NP2 in the ontology (Morin and Jacquemin, to appear). The knowledge needed for these methods mixes language-general information e.g. that natural language grammars can be usefully seen as deriving from a context-free base, and some language-specific information e.g. that “such as” tends to play the role described above in English.

Domain-specific methods utilize the existing ontology in conjunction with the terminological variation paradigm (Jacquemin, 2001). Ontological links can be inferred through systematic variations in syntax and morphology.

3.1. Domain General Aspects

The problem of ontology extraction lacks agreed principles and computational techniques, even more so than term
extraction. Although existing linguistic theory concerning
the effects of headedness, can be applied to the internal
structure of terms and provides predictions about their dis-
tributional properties, this is only possible because terms
are fundamentally linguistic objects. Ontology extraction,
contrast, concerns the underlying semantic structure of a
text, and the elements needed are only partially reflected in
surface structures. Both problem fields share the fact that
users can very easily determine whether a result is correct,
without necessarily being able to say anything about why.

Since ontology extraction is a semantic task by defi-
nition, it is tempting to apply techniques from formal sem-
tactics (e.g. (Pustejovsky, 1996)). However, while this
approach may be reasonable in principle, in practice it
rapidly becomes impractical. Not only are formal semantic
approaches typically limited to highly circumscribed do-
 mains, and effort intensive to extend, but they are invariably
computationally expensive. In contrast, project Parmenides
requires computationally straightforward methods in order
to work efficiently on large document collections, and must
be easy to tailor to different subject domains. For this ap-
lication, even regular parsing may be too intensive and in-
flexible, and we have begun with a template representation
approximately equivalent to regular expressions. There is,
in the following discussion, always a tension between lim-
itations on the expressive power of the formalism and the
need to express syntactic regularities useful for ontology
extraction.

The domain-general part of the TermFinder’s ontology
extraction component is based around a set of templates de-
signed to capture the syntactic signatures of basic ontolog-
cal constructions such as ISA, HAS_ A, HAS_PROPERTY,
and RELATED_TO.

We have defined a pattern language intended to make it
easy for users who are relatively linguistically sophisti-
cated, but unfamiliar with regular expression syntax to de-
fine template structures. TermFinder defines patterns over
feature structures containing lexical and part of speech
information for individual words, and allows a grouping op-
erator to fit element sequences into ontology templates. Since
the TermFinder provides a tagger, part of speech informa-
tion is always available. For example, (word=bank
tag=VB) is an instance of the word ‘bank’ used as a verb.
Alternation can also be used to specify the part of speech
tag as in NN|NNP. Both tag and word field are optional, al-
though if both are missing (all) must be used, a pattern that
matches any token. Regular expressions can be used di-
rectly, allowing the pattern (regex=-*ing tag=NN) to
match any noun ending in ‘ing’. Finally, the Kleene star
and plus operators work for whole elements, allowing (tag=NN)* to match zero or more nouns. A simple ex-
ample might be:

(1) (tag=JJ)* (tag=NN|NNP)+
 (word=is) (word=a)
 (tag=J)* (tag=NN)+
 1 ISA 2

Here the curly brackets are grouping operators that can
be referred to by number, defined by the order they appear in
the pattern. The first three lines define a pattern and the final

line assign its ontological interpretation. The template says
that when a noun group optionally prefixed by adjectives
follows the words ‘is’ and ‘a’, begins with a noun group
preceded by optional adjectives, the starting noun group
holds the relation ISA to the first. The grouping operator
ensures that surrounding adjectives are not included in the
ontological information.

The templates used in Parmenides are significantly
more articulated than this example, which would capture
only a small percentage of the ISA relations in text, but the
complexity necessary for real applications is considerably
eased by the availability of an intuitive feature representa-
tion.

Although the computational power of this representa-
tion is relatively weak, it has three significant advantages.
The first is that writing templates in this style is much more
intuitive than constructing regular expressions, even assum-
ing a way to provide the part of speech components were
found. The second advantage is that the templates can be
compiled down into finite state representations that are very
fast. And the third is that the style makes it relatively
straightforward to extend ontology extraction technology
into a partial parsing framework based on cascades of fi-
nite state transducers (Abney, 1996).

Rather than presenting the full range of templates and
results corresponding to each syntactic construction, we
focus the following discussion on the practical issues in-
volved in applying template structures to highlight the rela-
tionship between syntactic phenomena and the ontological
relation ISA, with particular focus on the strengths and lim-
itations of a weakly-expressive representation.

3.1.1. ISA relations

The example template 1, combined with easily gener-
ated variations can give surprisingly accurate performance
when extracting examples of the type:

(2) Snapple is a fortified juice smoothie
 NNP VBZ DT JJ NN NN

The approach begins to show its limitations when coor-
dination structures intervene because of the productive na-
ature of conjunction:

(3) meal replacements are a reliable and safe
 NN NN VBZ DT JJ CC JJ
 method of dieting
 NN IN NN

Intervening material is a general issue with the tem-
plating system, and clearly suggests a parsing framework.
However, in practice the infrequency of conjoined ‘is
a’ constructions often does not justify a more complex
method.

The sequence ‘such as’ can be a surprisingly accurate
indicator of ISA information, using templates of the form:

(4) (tag=JJ)* (tag=NN|NNS)+ (word=such) (word=as)
 (tag=JJ)* (tag=NN|NNS)+
 2 ISA 1
This matches the initial noun phrase, and first conjunct (and with straightforward augmentation, also the other conjuncts) of sentence 5 giving an accurate report of their type relation. Templates constructed this way tend to miss property structure when there are intervening qualifiers, usually prepositional phrases. In sentence 6, for example, ‘consumption’ qualifies the intended target, and will tend to be matched instead of ‘fat’.

(5) basic foods such as wheat and soya
   JJ NNS JJ IN NN CC NN

(6) dietary properties such as dietary stability
   JJ NNS JJ IN JJ NN
   and consumption of fat
   CC NN IN NN

(7) DNA binding in BCP is a result of
   NN VBG IN NN VBZ DT NN IN
   IL-7
   NN

There are several problems here: First, a naive template will make ‘BCP’ a type of result. Second, if the template is altered to allow VBG as a type then ‘in BCP’, a prepositional phrase, needs to be ignored. More problematically, allowing verb phrases to be subclass targets tends to generate a large number of spurious matches involving verbs from previous clauses.

The syntactic complexity and wide range of distributional profiles of natural language expressions of events and processes makes them particularly hard to spot, even in cases where the presence of a word sequence is in fact, a correct indicator that a relation exists. ‘Such as’ constructions containing VBG headed con- or disjuncts (8) will often fail on the template approach due to the relative lack of part of speech and position constraints placed on daughters by a verb, compared to those imposed by a head noun. Fortunately, verb headed disjuncts like 8 appear to be a minority.

(8) process changes such as reducing the
   NN NNS JJ IN VBG DT
   amount of magnesium added or by adding
   NN IN NN VBD CC IN VBG
   ascorbic acid
   JJ NN

More explicitly ontological issues arise when the first argument, rather than the conjuncts is qualified. In sentence 9 the difficulty is to distinguish that the prepositional phrase modifying the initial noun group is the target superclass, not the nearer noun group. Although a syntactic treatment would deal with this form of qualification naturally, the effect of ignoring it is also small in practice.

(9) Methods in the prevention of heart decese
   NN IN DT NN IN NN NN
   , such as arginine consumption
   , JJ IN NN NN

While this discussion has detailed the limitations of a template-driven approach to domain-general ontology extraction, mostly due to limited expressive power, it is striking that the constructions that are problematic in theory are a minority of instances in Parmenides applications; possible ontologically informative constructions far outnumber probable ones. This is in a large part due to our emphasis on research and development materials (particularly experimental reports, product descriptions) which typically use restricted forms tailored to the effective communication of information. Limitations that are not due to the expressive power of the pattern language can be addressed by making use of domain-specific ontological resources, to which we now turn.

3.2. Events and Processes

Parmenides is concerned not only with the extraction of ontological relationships between objects, but also between processes and events, and possibly causation. These structures constitute a challenging problem for ontology extraction, and one that arises often. An example of the difficulty with using ‘is a’ is:

There are several problems here: First, a naive template approach due to the relative lack of part of speech that operate on the Head of a MWT and those that operate on the Modifiers.

The most obvious ontological structure results from modifier expansions. For example, ‘iron absorption’ → ‘deficient iron absorption’ indicates an ISA relation as the two MWTs share a common head. This example is clear as the shorter MWT remains unchanged in the longer MWT. However, there is less certainty when the shorter MWT is changed, ‘iron absorption’ → ‘iron mineral absorption’. In this example (sometimes called an insertion) the internal structure of the shorter MWT has changed rather than simply being added to. With this in mind, Modifier Expansions are exploited to produce ISA hierarchies across the extracted terminology.

The application of Head Expansions is not an straightforward as they do not consistently result in a definable ontological relation. For example, how is ‘ascorbic acid’ related to ‘absorbic acid metabolism’? Intuitively, there is an
involvement relation as the second MWT clearly ‘involves’ the first but this difficult to formalize and always assuming a relation is unreliable.

Similarly, unconstrained substitution only results in a specific relation on a fairly hit and miss basis. For example, ‘absorbic acid intake’ and ‘absorbic acid consumption’ are clearly related semantically but the same substitution variation also identifies ‘absorbic acid test’ - a less relevant MWT. By further constraining the substitution relation to hold only between tokens already linked in the ontology such spurious matches are eliminated. This process identifies three types of SEE-ALSO relations when the ontological link between substituted tokens is synonymy (Dowdall et al., 2003), (Hamon and Nazarenko, 2001):

- Strong - head substitution: ‘normal human’, ‘normal person’, ‘normal individual’
- Intermediate - modifier substitution: ‘gender difference’, ‘sex difference’
- Weak - head and modifier substitution: ‘hormone effect’, ‘endocrine event’

Additionally, head substitutions identify ISA relations when the substituted tokens are already defined in the ontology as hyper/hyponyms.

So specific expansions and substitutions result in easily definable ontological links. As for the rest, the lack of correlation between variation and semantic link makes them unsuitable for ontology expansion. However, to ignore them is to ignore the linguistic patterns that that exist across the concepts of a domain. Head Expansions and substitutions which are not ontologically linked are useful during query formulation. If a query concept does not appear in the domain the graceful fall back is to suggest ontological relations when the substituted tokens are already defined in the ontology.

5. Conclusions

Parmenides integrates information, MWT and ontology extraction methodologies within a single knowledge management framework. One of its most linguistically interesting aspects is therefore that it uses approximately the same methods (templatting, named-entity recognition, shallow parsing) to generate ontological structure, as it uses to perform ontology-backed information extraction. This promises to make the gap between MWT finding, ontology construction and ontology use much smaller than it usually is. When an ontology is constructed offline by topic experts, there is no guarantee that the distinctions it contains can actually be found in raw text at all. Parmenides’ approach keeps the two aspects of use and construction synchronised by obliging them to depend on mostly the same methods.

Once extracted MWTs provide a point of access into the domain and reveal candidate ontological links that exist in the corpus, as well as the links between MWTs and the existing ontology.

6. References

Abney, Steven, 1996. Tagging and partial parsing. In Ken Church, Steve Young, and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech*. Dordrecht: Kluwer Academic Publishers.

Castelvi, M. T. C., R. E. Bagot, and J. V. Palatresi, 2001. Automatic term detection: A review of current systems. In *Recent Advances in Computational Terminology*. John Benjamins, pages 53–88.

Dowdall, J., F. Rinaldi, F. Ibekwe-SanJuan, and E. SanJuan, 2003. Complex structuring of term variants for question answering. In *Proceedings of the ACL Workshop, Multi-Word Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan.

Frantz, K. and S. Ananiadou, 1996. Automatic term recognition using contextual clues. In *Proceedings of Mulsaic 97, IJCAI*.

Hamon, T. and A. Nazarenko, 2001. Detection of synonymy links between terms: Experiment and results. In *Recent Advances in Computational Terminology*. John Benjamins, pages 185–208.

Jacquemin, C., 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Kageura, K., 2002. *The Dynamics of Terminology: A descriptive theory of term formation and terminological growth*. John Benjamins.

Morin, E. and C. Jacquemin, to appear. Automatic acquisition and expansion of hyponym links. *Computer and Humanities*.

Pustejovsky, James, 1996. *The Generative Lexicon*. Cambridge MA: MIT Press.

Taghva, K. and J. Gilbreth, 1999. Recognizing acronyms and their definitions. *In the International Journal on Document Analysis and Recognition (IJDAR)*, 1:191–198.