

Enhancing Knowledge Mapping Using Automatically Derived Concepts

Anjo Anjewierden
Human Computer Studies Lab.
University of Amsterdam
Kruislaan 419, Amsterdam
anjo@science.uva.nl

Willem-Olaf Huijsen
Telematica Instituut
PO Box 589
7500 AN Enschede
Wolf.Huijsen@telin.nl

Marjan Grootveld
Telematica Instituut
PO Box 589
7500 AN Enschede
Marjan.Grootveld@telin.nl

ABSTRACT

Knowledge-mapping tools enable users to quickly identify relevant information and expertise. This paper discusses a number of natural-language phenomena that limit the performance of a straightforward approach. An empirical study on a real-life community provides a quantitative indication of the impact of this noise on the markup of concepts and retrieval of documents. We then discuss enhancing the usefulness of knowledge-mapping tools through automatically derived concepts.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Text analysis

1. INTRODUCTION

This paper addresses the problem of supporting the development of ontologies from corpora of text messages created by Communities of Practice (CoPs) with the objective of metadating, markup and retrieval. We call these knowledge-mapping tools. We describe the application of the tools and techniques (in a system called TOKO) in a practical real-life context: The Extrusion Reliability Community (ERC) at Basell.

2. NOISE REDUCTION

The objective of noise reduction is to identify as many lexical and syntactical variants as possible for the set of concepts in an ontology given a document base. We describe the different kinds of noise that occur and hint at techniques of compensating for them.

Alternate spellings and inflections. The dictionary we use, CELEX, contains the alternate spellings and inflections of common words. **Misspellings.** These are

relatively frequent in the ERC. Our detection is based on the notion of Levenshtein (edit) distance [1], the number of character transformations required to obtain a correctly spelled word. **Compound contraction.** The orthography of new, domain specific compound terms is highly dependent on the language. In English, it is common to keep words separated, e.g. “domain name”, whereas Dutch and German contract constituent words “domainname”. Compounds are easy to recognise simply by inserting or removing the word separator and checking whether the contracted variant occurs in addition to the terms mentioned in the ontology. **Abbreviations.** Abbreviations are extremely frequent in the ERC. One of the most promising algorithms has been suggested by Schwartz and Hearst [3]. They assume that the letters in the short form appear in the long form in the same order and that the short and long forms appear next to each other and one of the forms is bracketed. Unfortunately, and not surprisingly, ERC messages do not have such a consistent style. **Synonyms.** Enumerating synonyms is beyond the capabilities of any software that is not even present in the documents. TOKO implements several algorithms to suggest synonyms, or more generally semantically similar terms. The algorithm based on considering the immediate lexical context before and after a term works best. In practice, synonyms are added to the ontology by the knowledge engineer. **Conflations.** We define a conflation [2] to be a syntactic paraphrase of a reference to a concept. [knife grinding] can appear in a document as the conflation “grind the knives” and [filter cleaning] can be conflated to “filters are cleaned”. We derive syntactic rules to find conflations are based on an algorithm that operates automatically: (1) Generate all possible word orders of compound terms in the ontology and derive a pattern by inserting a **near(2)** operator (which matches 0, 1 or 2 arbitrary words) between these words. [extruder problem] results in the patterns {extruder near(2) problem} and {problem near(2) extruder}. (2) The patterns are run on the corpus and all matches that contain nouns, e.g. “extruder screw problem”, or verbs also in the ontology are deleted. (3) Re-

Ontology	Markup		Retrieval	
	Count	Percentage	Count	Percentage
Ontology	19699		10942	
Misspellings (M-)	-133	0.6%	-75	0.6%
Conflations (X-)	-194	0.9%	-162	1.4%
Contractions (C-)	-797	4.0%	-250	2.2%
Synonyms (S-)	-1259	6.3%	-619	5.6%
Abbreviations (A-)	-1407	7.1%	-666	6.0%
M-C-	-930	4.7%	-326	2.9%
A-M-C	-2337	11.8%	-996	9.1%
S-A-M-C	-3600	18.2%	-1628	14.8%
X-S-A-M-C	-3794	19.2%	-1790	16.3%

Table 1: Concepts marked up and documents retrieved for all concepts in the ontology.

Ontology	Markup		Retrieval	
	Count	Percentage	Count	Percentage
Ontology	2283		1656	
Abbreviations (A-)	-0	0.0%	-0	0.0%
Misspellings (M-)	-2	0.0%	-2	0.0%
Synonyms (S-)	-82	3.5%	-60	3.6%
Contractions (C-)	-217	9.5%	-125	7.5%
Conflations (X-)	-194	8.4%	-162	9.7%

Table 2: Markup and retrieval for compounds.

maining matches are generalised by replacing the concept words by a wildcard for any word. The conflation “problems with the extruder” results in the conflation pattern {• with the •}. In the ERC forum, 70 conflation patterns were found and manual inspection indicates most are relevant.

Here we provide a quantitative empirical study of the impact of correcting for noise. We assume the user is interested in both markup and retrieval. For markup it is necessary to identify all occurrences of a concept in a document. For retrieval it is only necessary to find at least one occurrence of a concept in a document. An example of the distinction is “PP” (an abbreviation for [polypropylene]). For retrieval, efforts to define the abbreviation are wasted iff every document that contains the abbreviation also contains the long form. The experiment was conducted on the ERC forum (1830 documents, 756 concepts in the ontology). Table 1 summarises the results. The first line contains the total number of concepts marked up and documents retrieved, utilising all noise reduction techniques and represents an upperbound on associating the content of the forum with the concepts in the ontology. The remaining lines show the impact of dropping the recognition of variants. The second experiment conducted was to make a distinction between *simple* domain terms and more complex notions. There is little incentive for the end-user to inspect all documents containing [extruder] (486). It is much more attractive to formulate a more

specific query related to a particular issue at hand: “extruder failure”. Table 2 summarises the results. Conflations are now significant. 9.7% of the documents would not be retrieved if we ignore them. Although the results of the experiments are ontology and the forum specific we now have at least some empirical evidence that noise affects retrieval, and that compensating for noise is worth considering in real-life contexts.

3. DERIVED CONCEPTS

Inspired by the second experiment we developed a means to automatically derive complex concepts from basic concepts. This makes the retrieval mechanism understands the query as a single concept and not just as the co-occurrence of the words that make up the query. Consider the query “die plate cleaning”. [die plate] is a [component] in the ontology and [cleaning] is a [task]. A search for all documents that contain both [die plate] **and** [cleaning] returns 17 documents. A search for [die plate cleaning] as a single concept returns 6 documents including conflations. This suggests that [die plate cleaning] as a concept results in better precision, and should therefore be included in the ontology. Manual inclusion is, however, impractical. 60 components and 40 tasks in the ontology yield 2400 derived concepts to consider. Our implementation derives concepts automatically based on the lexical variations and the conflation patterns. We first run the markup algorithm and then retrieve the derived concepts by matching the conflation patterns. If we start the search with [cleaning] it suggests derived concepts based on components that can be cleaned: [die plate], [start-up valve], [silo] etc. Starting from [die plate] we obtain the tasks [grind], [replace], [repair], [analyse] etc.

4. CONCLUSIONS

We discussed a range of natural-language phenomena that impede a straightforward markup of concepts in texts. An empirical study analysing the content of a real-life, technical Community of Practice, suggests it is necessary to take these “noise” phenomena into account. We also showed that the same phenomena can be used to derive concepts dynamically and automatically given a set of (simple) base concepts.

5. REFERENCES

- [1] V. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Russian Problemy Predachi Informativii*, 1:12–25, 1965.
- [2] A. Savary and C. Jacquemin. Reducing information variation in text. In S. Renals and G. Grefenstette, editors, *Text- and Speech-Triggered Information Access*, volume 2705, pages 145–181. 2003.
- [3] A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings Pacific Symp. Biocomputing*, Kauai, Hawaii, January 2003.