Concept Maps: Theory, Methodology, Technology
Proc. of the Second Int. Conference on Concept Mapping
San José, Costa Rica, 2006

AUTOMATIC CREATION AND TRANSLATION OF CONCEPT MAPS FOR COMPUTER SCIENCE-RELATED THESES AND DISSERTATIONS

Ryan Richardson, Ben Goertzel, Edward A. Fox, Virginia Tech, USA
Hugo Pinto, VettaLabs, Brazil
Email: ryanr@vt.edu, ben@goertzel.org, fox@vt.edu, hugo@vettalabs.com

Abstract. Concept maps are often used as tools to enhance student learning on new topics. We hypothesize that they also can be used as summarization tools for large documents. We hypothesize further that, because concept maps usually have short text sections (words and short phrases), it should be easier to translate concept maps than parts of documents (such as abstracts), since the text need not be natural-sounding sentences. We have selected electronic theses and dissertations (ETDs) as our sample collection, since we have large numbers of these in both English and Spanish. We have adapted a tool that automatically produces semantic maps from text, called Relex, to produce concept maps of ETDs for computing documents, using an ontology of computer science. We also have developed code to automatically find translations for English phrases by using a Spanish ETD collection as a corpus.

1 Introduction

The growth of the World Wide Web has led to increased availability of many large documents, such as electronic theses and dissertations (ETDs). However, it is difficult for users to determine if such a large document, written in a language they cannot read, is relevant to their information needs—so they can seek a translation of it, or arrange to have it translated. Unfortunately, automatically translating large documents, like ETDs, so they easily can be found, read, and understood, is beyond the current state of the art in machine translation, especially in technical fields.

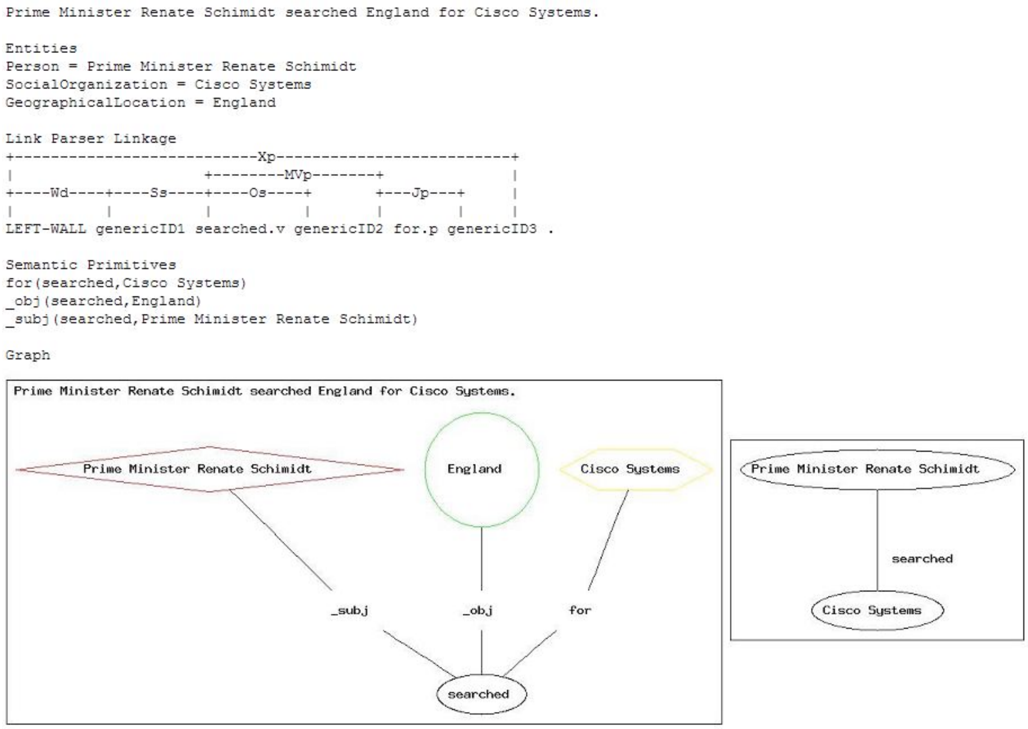
2 Automatic Generation of Concept Maps

Since concept maps consist of nodes and relations that are stated clearly and succinctly, we hypothesize that concept maps are more readily translatable than full-length text. When translating ordinary text, much care must go into assuring that one obtains clear and natural-sounding sentences. This is not such a big issue in concept maps, which tend to have text that is single words, short phrases, or short sentence fragments. We further hypothesize that, if there were a way to automatically produce a concept map that gives a reasonable view of the content of a document, we could then translate the concept map automatically, thus yielding a summary of the document in the new language (Richardson and Fox 2005). This would serve as an *indicative* summary, helping information seekers decide whether or not they should try to have the original document translated into a language that they can read.

2.1 The Relex Tool

Relex is a system that translates syntactic dependencies into a graph of semantic primitives (Wierzbicka 2006), by means of template matching algorithms. Relex also detects implied quantities, normalizes passive and active forms into the same representation and assigns tense and number to sentence parts. Relex must be used in conjunction with a dependency grammar parser; in the current Relex version, the parser used to extract the base dependencies is CMU’s link parser (Sleator and Temperley 1993). Most of Relex’s morphological analysis is performed using WordNet's morphological functions (Fellbaum 1998).

The link parser often breaks when presented with unknown, multi-word entities, such as “Prime Minister Gerhard Schroder” or “Cisco Systems Inc.”. Instead of adapting the link parser to better handle such phenomena, we have taken an alternate approach, in which we tag named entities and replace them with simple identifiers. For instance, instead of feeding the sentence “Donald Macloud Jr. is going to Belo Horizonte to assume the post of dean at Abraxas University”, the system feeds “ID1 is going to ID2 to assume the post of dean of ID3” into the parser. All substituted entities IDs are treated as noun phrases. After the sentence is parsed and converted into a semantic primitive graph, the IDs are restored back to full entity names, and each entity is assigned its proper tag (in the example above, Person, Location, and Organization, respectively). Relex comes with a set of tools to extract and draw paths between named entities in the semantic primitive graph. Figure 1 shows the link parser output, semantic primitive based relationships, and inter-entity paths resulting from Relex analysis of the sentence “Prime Minister Renate Schimidt searched England for Cisco Systems”.



```
Prime Minister Renate Schimidt searched England for Cisco Systems.

Entities
Person = Prime Minister Renate Schimidt
SocialOrganization = Cisco Systems
GeographicalLocation = England

Link Parser Linkage
+--------------------------Xp---------------------------+
|                     +--------MVp-------+              |
+----Wd----+----Ss----+----Os----+       +---Jp---+     |
|          |          |          |       |        |     |
LEFT-WALL genericID1 searched.v genericID2 for.p genericID3 .

Semantic Primitives
for(searched,Cisco Systems)
_obj(searched,England)
_subj(searched,Prime Minister Renate Schimidt)

Graph
```

Prime Minister Renate Schimidt searched England for Cisco Systems.
Prime Minister Renate Schimidt
England
Cisco Systems
_subj
_obj
for
searched
Prime Minister Renate Schimidt
searched
Cisco Systems

Figure 1. Various stages of Relex Processing Pipeline**.**

Relex is implemented within IBM’s UIMA framework (IBM 2006). This framework describes a series of design patterns, interfaces, and metadata to implement, combine, and deploy analysis capabilities. The default entity tagger used with Relex is based on the Another Nearly-New Information Extraction (ANNIE) system, distributed as part of the General Architecture for Text Engineering (Cunningham, et al. 2002). This tagger was customized to better deal with multi-word and non-English names.

2.2 An Ontology for Computing and Information Sciences

Relex comes with two customized entity taggers, one specialized for the biomedical domain and another for world news and finance. Neither of these is appropriate for our purposes in the current investigation. In order for Relex to be able to recognize computing terms as entities, we needed a comprehensive (and consistent) source of terminology. We found this in a project called the Ontology Project (Cassel 2006), in development at Villanova University and sponsored by ACM. This project has divided computing into 21 topic level domains, and provides a hierarchy of terms that go from one to six levels deep for each of these main topic areas. The ANNIE extractor needs to recognize when an instance of a class in the ontology appears in the text. We accomplished this by selecting various nodes in the ontology and encoding “gazetteers” – lists of individual terms that would be the surface representations of that node in the document. Since the computing ontology has about 900 leaf nodes, it would be very time-consuming to write gazetteers for all of them. Therefore we selected a few areas of computer science for which Virginia Tech has a large number of ETDs (i.e., digital libraries, human-computer interaction, virtual environments) and wrote gazetteers for only these (200+) nodes.

2.3 Generation of the Map

We developed a script that separates an ETD into chapters, which are passed to Relex. Relex processes one chapter at a time, and produces a concept map (see **Figure 2**). The map shows the identified concepts as nodes, and relationships between them as links, labeled with the most common verbs that identify that relation. **Figure 2** illustrates a concept map for a chapter selected from an ETD. Since we believe that our approach can work for any domain, not just computing, we have picked one ETD dealing with current events.

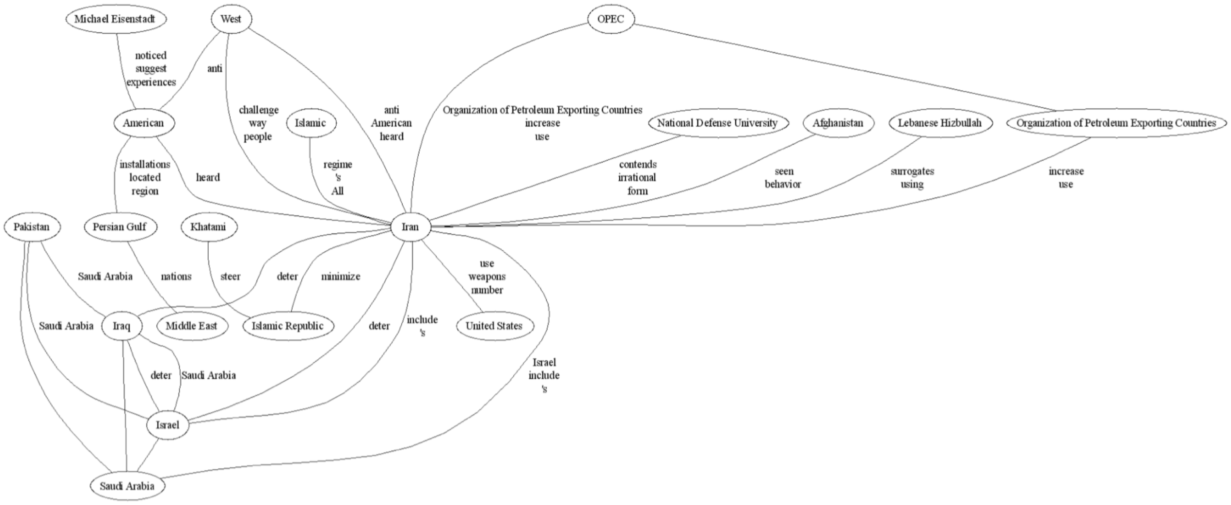


Figure 2. A Relex-generated concept map of one chapter of the Virginia Tech ETD "The Nuclearization Of Iran: Motivations, Intentions and America's Responses" by John Hanna.

3 Automatic Translation of Concept Maps

Since Relex currently only works for English documents, we are limited to producing English maps, which we then translate into Spanish. We have obtained a collection of English and Spanish ETDs from various sources. We have over 200 computing ETDs (in English) from the Virginia Tech collection. Our Spanish computer science collection consists of about 100 ETDs from the Universidad de las Américas (UDLA), 417 from other NDLTD members (all from universities in Spain), and 78 ETDs about computing through an arrangement with the Universidad Nacional Autónoma de México (UNAM).

Using an algorithm described in (López-Ostenero, et al. 2004), we have been able to mine these comparable English/Spanish corpora to find translations for technical phrases that would not appear in any freely-available dictionary. We have enhanced López-Osteñero's algorithm in 3 ways:

1. López-Osteñero's algorithm only translated 2 or 3 word phrases. Our implementation will translate 2, 3, 4, or 5 word phrases, since phrases with these lengths are common in concept maps.
2. López-Osteñero's algorithm will only translate an n word phrase to another n word phrase. Our implementation will translate a phrase of any length to one of any other length.
3. In the case where the first word in the source phrase has many possible translations into the target language, there will be a huge number of candidate phrases to analyze. Our implementation finds the word with the fewest translations into the target language, and swaps the first word with the word with fewest translations. This does not affect translation quality and greatly increases speed of processing.

English Phrase	**Mined Spanish Translation**
Bibliographic information	información en término bibliográfico
collision detection	detección de colisión
Color channel	canal de color
flat rendering	representación del plano
information visualization	visualización de información
object recognition	reconocimiento de objeto
Quality of service	calidad de servicio
query formulation	formulación de pregunta
ray tracing	trazado de rayo
visualization hardware	dispositivo de visualización

Table 1. Examples of English technical phrases and the translations that were mined from our Spanish ETD collection.

Currently we are using a 45,000 word list, which was mined at the University of Maryland from parallel corpora, to translate individual words. In cases where a word in an English concept does not have a translation in our bilingual word list, we use BabelFish (Systransoft 2006) to translate it. **Table 1** gives examples of translations of technical phrases from English to Spanish that we have mined from our Spanish ETD collection.

4 Proposed User Study

We are in the final planning stages of an experiment to test the effectiveness of automatically-generated and translated concept maps as summarization aids for large documents. The subjects will be Spanish speakers who are computer science majors at UDLA, and who will participate in the experiment via the Web. For the experiment, users will be randomly assigned into 3 groups, corresponding to 3 conditions. The three conditions will be 1) having access to the abstract of an ETD, 2) having access to an automatically generated concept map of an ETD, and 3) having access to both. No group will have access to the original English ETD, to avoid the problem that some UDLA students will be better at reading English than others. This also simulates the common situation where users can access an ETD's abstract, but not the full text, for copyright or technical reasons (for instance when the ETD exists only on paper). The concept maps will be translated via our implementation of López-Osteñero's algorithm, with single words being translated using the previously-mentioned University of Maryland word list. For the abstract condition, the abstracts will be translated into Spanish via Babelfish. We plan to present users with 5 information seeking tasks. The website code will randomize the presentation of the 5 tasks to the subjects. For each task, they will be given an abstract, a concept map, or both (depending on their group), taken from about 50 ETDs, and asked to determine which ETDs are relevant to a particular research question and which are not. The relevant documents will be determined beforehand by a group of 3-4 experts. The relevance determinations of the students will be compared to that of the experts.

We hypothesize that Spanish speakers with access to concept maps (either with or without abstracts) will perform better in relevance determination than those who do not have access to concept maps. By the time of the conference, we hope to have solid confirmation of our hypothesis, and to be able to present examples and statistical analysis results from the experiment.

5 Acknowledgements

This Research Project was supported by the RDEC/ACE project. Funding from IBM will assist with work with UIMA. We thank Vicente Cordeiro at VettaLabs and Michael Ross at SAIC for their support and work on Relex. We also thank Lillian Cassel and the team at Villanova University for their work on the ontology of computing and information technology, and ACM for their support. We thank Silvia Marín and Alberto Castro Thompson for giving us access to UNAM's ETD collection. Finally,we thank Alfredo Sanchez for allowing his students to participate in the cross-lingual experiment at UDLA.

6 References

Cassel, L. N. (2006). "The Ontology Project." Retrieved 19 April, 2006, from http://what.csc.villanova.edu/twiki/bin/view/Main/OntologyProject.

Cunningham, H., D. Maynard, et al. (2002). *Framework and Graphical Development Environment for Robust NLP Tools and Applications*. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, Massachusetts, The MIT Press.

IBM. (2006). "Unstructured Information Management Architecture (UIMA)." Retrieved June 6, 2006, from http://www.alphaworks.ibm.com/tech/uima.

López-Ostenero, F., J. Gonzalo, et al. (2004). "Noun Phrases as building blocks for Cross-Language Search Assistance." *Information Process and Management* **41**: 549-568.

Richardson, R. and E. A. Fox (2005). *Evaluating Concept Maps As A Cross-Language Knowledge Discovery Tool for NDLTD*. In Proceedings of Electronic Theses and Dissertations Conference, Sydney.

Sleator, D. and D. Temperley (1993). *Parsing English with a link grammar*. Third International Workshop on Parsing Technologies, Tilburg, Netherlands & Durbuy, Belgium.

Systransoft (2006). Babelfish, Systransoft.

Wierzbicka, A. (2006). Semantics, Primes and Universals. Oxford, UK, Oxford University Press.