AUTOMATIC DISCOVERY OF WORD SEMANTIC RELATIONS

Gaël Dias, Rumen Moraliyski, João Cordeiro, Antoine Doucet and Helena Ahonen-Myka

Abstract. *In this paper, we propose an unsupervised methodology to automatically discover pairs of semantically related words by highlighting their local environment and evaluating their semantic similarity in local and global semantic spaces. This proposal differs from previous research as it tries to take the best of two different methodologies i.e. semantic space models and information extraction models. It can be applied to extract close semantic relations, it limits the search space and it is unsupervised.*

Keywords: Synonymy, language acquisition
2010 Mathematics Subject Classification: 68T50

1. Introduction

Thesauri, that list the most salient semantic relations between words have mostly been compiled manually. Therefore, the inclusion of an entry depends on the subjective decision of the lexicographer. Consider the following simple sentence: *The words of a phrase relate in many ways to each other.* Probably only the pair <*word , phrase*> would be listed in a manual resource with its semantic relation, but interpretation clues for a polysemous word like *way* would be more difficult to code in a thesaurus. In text understanding, humans are capable, up to a variable extent, of uncovering those relations. Natural language processing systems, however, need either a complete inventory of the semantic relations or a module able to infer them from the text in order to perform human like interpretation.

Numerous attempts in automatic thesaurus construction are known [6]. The entries that they extract comprise long lists of terms related to the head in unspecified ways. An attempt to partially annotate such thesauri with semantic information is made in [14]. Apparently, applying various classifiers and filters consecutively improves the precision at cost of recall.

Other works, make use of exhaustive search over the vocabulary to induce semantic relations [10]. The exhaustive search is the obvious way to verify all the possible connections between words of the vocabulary. However, comparison based on word usage can only highlight those terms that are *highly* similar in meaning [16]. Thus, the exhaustive search is only capable of finding the most salient semantic relations. At the same time, neologisms, recently adopted foreign words and names, which consist that part of the current vocabulary that needs constant update, elude characterization.

To overcome the difficulties encountered by exhaustive global search in semantic spaces, some works propose to exploit local patterns to extract hyponymy [9], synonymy [3] or meronymy relations [2]. These works either study manual patterns or propose to automatically acquire relevant local patterns based on supervised learning. Although, these approaches present successful results they still require manually annotated data for training.

In order to discover pairs of semantically related words, we need to have them highlighted by their environment. For this purpose we first extract pairs of paraphrases from automatic clusters of Web news stories. Then we align the paraphrases in order to highlight the equal and different parts and create Test of English as a Foreign Language (TOEFL) -like tests of one target word plus an as short as possible list of words that are predominantly in paradigmatic relations with the target.[1] Eventually, we introduce a contextual similarity measure and an characterization of its ability to highlight close semantic relations.

2. Related Work

Many early attempts to discover related words in an automated manner employed some syntactic parser in order to avoid irrelevant processing [11]. However, a common problem encountered was the limited size of the available corpus and for this reason they used to focus on restricted domains. In a more recent work [10] shows that the behavior of contextual similarity measures depends on frequency but as well on semantic specificity and semantic classes of words. Although strong conclusions can not be drawn, since comparison with the corresponding WordNet quantities is missing, it is still apparent that contextual similarity measures have a tendency to detect semantic relations beyond mere synonymy.

Most of the known methodologies compile thesaurus entries from lists of words related to the head in unspecified ways. Here emerge two diverging families of work. One is established by [13] with the introduction of the synonymy part of the TOEFL as an evaluation problem for synonymy discovery techniques. The other direction takes a more general view and aims at automatic annotation of existing lexicons with semantic information or at building resources listing only pairs of words in some specific semantic relation. [9] first describes a work where patterns that convey the desired semantic relation are manually selected and subsequently pairs of words that fit the patterns are gathered.

In order to avoid exhaustive search and manual annotation of training data we take the most of both strategies i.e. the one looking at common

[1] Paradigmatic are those relations that factorize a set in classes of interchangeable units. Here we want to automatically discover paradigmatic classes of semantically related words - synonyms, hypernyms, hyponyms.

patterns and the other one using distributional semantics analysis. In the next sections we give motivation, technical details and evaluation of this process.

3. Creating Test Cases

The Distributional Hypothesis, formulated in [8], suggests that counting the contexts that two words share improves the chance for correct guessing whether they express the same meaning or not. The plausibility of this assumption is supported by the psycho-linguistic research [12, 16] and numerous empirical studies.

It takes three steps to create automatically TOEFL-like test. First we employ an algorithm for paraphrase extraction, detailed in [4]. Paraphrase is a restatement of a text or passage, using other words. This is often accomplished by replacing words with their synonyms, hyponyms or hypernyms and changing word order. For example the sentences in Figure 1, taken from web news stories excerpts, are paraphrases of a news about the release of a comic movie and show that "*feature*" is possibly substituted by *"news"*, *"controversy"*, *"comedy"* or *"film"* and as such may share common meanings.

(1) *Kazakhs are outraged by the wildly anticipated mock documentary feature Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*

(2) *The news follows controversy surrounding the comedy film Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan which cut so close to the funny bone.*

Figure 1. A sample of 2 paraphrases.

It is desirable to identify paraphrases which have certain level of dissimilarity, because this is precisely what will open room for semantic relation discovery. The method developed in [4] satisfies this requirements as it was developed with lexical acquisition task in mind.

Next we align the paraphrases applying the algorithm given in [1]. The longest possible non contiguous sequence of words common for both paraphrases is found and the subsequences between the common parts are the differences where the useful substitutions occur.

From our example in Figure 1 the longest common sequence is "*Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan*". Due to its sequential property, passing in parallel through the paraphrases, we are bound to encounter the word "*Borat:*", and any word encountered before is not common for both sentences. Once "*Borat:*" is encountered, we know that we are bound to encounter the second word of the common sequence, "*Cultural*". Thus, in only one pass, we can obtain the alignment presented in Figure 2.

> [{1:*Kazakhs are outraged by the wildly anticipated mock documentary feature*} {2:*The news follows controversy surrounding the comedy film*}] Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan [{2:*which cut so close to the funny bone*}]

Figure 2. The alignment corresponding to the sentences of Figure 1. Word sequences without brackets are common to both sentences. The sequences in curly brackets are specific to the sentences with the corresponding numbers.

The final step is to form TOEFL-like test cases from the aligned segments. The notion of test implies one word in a specific position for which we are searching matches among a list of candidates. Those parts of the paraphrases that lie between two successive parts of the common sequence have different orthographic appearance, nevertheless, we assume that they have similar meanings since they are both parts of paraphrase sentences and share left and right contexts. Therefore, here is the place where we search for word substitutions. As we are interested in nominal semantic relations and only open class words with the same part-of-speech are eligible candidates we lemmatize and assign part-of-speech tags to the aligned paraphrases using MontyLingua [15]. Precisely, the construction of the tests goes like the following algorithm.

```
For each aligned sub-segment
    For each open class word
        Create a list of candidates from
        the rest of the segments that share
        left and right contexts.
    End
End
```

From the aligned paraphrase in Figure 2 we extract two test cases for the target words "*kazakh*" and "*feature*" as shown in Figure 3.

(1) *kazakh* | *news* | *controversy* | *film*
(2) *feature* | *news* | *controversy* | *film*

Figure 3. Two TOEFL like test cases.

4. Measuring Similarity between Words

Now as we have the TOEFL-like cases, we need a method to select the best candidate. In this context, we must evaluate the similarity between two nouns which are represented by their respective word context vectors $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip})$ of observations on p variables (or attributes). The similarity between two vectors X_i and X_j is defined as $S_{ij} = f(X_i, X_j)$ where f is a function of the observed values.

For our purpose, the attributional representation of a noun consists of tuples $\langle v, r \rangle$ where r is an object or subject relation and v is a given verb appearing within this relation with the target noun.

In order to mitigate the effect of very frequent and non-informative contexts and the noise of very infrequent contexts we calculate a value of association between the attribute $\langle v, r \rangle$ and the characterized word n, called Pointwise Mutual Information (PMI) [17] as defined in Equation 1. We estimated the probabilities based on a Web corpus of 500 million words.

$$(1) \qquad PMI(\langle n|r \rangle, \langle v|r \rangle) = \log_2 \frac{P(n, v|r)}{P(n|r)P(v|r)}$$

To quantify the similarity between two context vectors, the Cosine similarity measure is usually applied and estimates to what extent two vectors point along the same direction. It is defined in Equation 2.

$$(2) \qquad cos(n_1, n_2) = \frac{\sum_{k=1}^{p} n_{1k} n_{2k}}{\sqrt{\sum_{k=1}^{p} {n_{1k}}^2} \sqrt{\sum_{k=1}^{p} {n_{2k}}^2}}$$

5. Results and Discussion

In this section, we evaluate our methodology over a set of web news. This environment proves to be very fruitful for paraphrase extraction, since many sentences convey the same message but in a varied form. We gathered 3 days of news from the Google News website[2]. From these texts were extracted 27 thousand pairs of paraphrase sentences. After alignment this resulted in a set of 22 thousand TOEFL like test cases with an average of 4.6 candidates.

5.1. TOEFL like tests. In order to keep the evaluation manageable, we retained at random 1000 noun test cases, which we manually classified into 5 classes with respect to whether the test contained a pair of words in one of the following relations: *Synonymy*, *Co-Hyponymy*, *Is-a*, *Instance-of*. Otherwise, we labeled it as *None*. Only afterwards we used the test set for evaluation as to avoid any influence of the automatic classification on the manual labeling. In Table 1, we present the distribution of the tests per category.

Table 1. Classification of the Test Cases.

Synonyms	**Co-Hyponyms**	**Is-A**	**Instance Of**	**None**	**Overall**
117	108	61	86	628	1000

[2] http://news.google.com/

About 35% of the wrong test cases are accounted for by bad preprocessing i.e. incomplete removal of XML and HTML tags and wrong sentence retrieval. Significant reordering impedes the algorithm to find correct alignment. In these cases the alignment is anchored around very common, uninformative words. Only occasionally such alignments result in sensible tests. Other details, as for example better part-of-speech tagging, anaphora resolution and named entity recognition, need to be fulfilled in order robust paraphrase extraction and alignment to be achieved. See [7] for a more detiled error analysis.

Table 2. Manually annotated tests. The respective relations hold between the first and the second words of each test.

Synonyms:	*body* | *panel*
	michael | *mike*
	administration | *government*
Co-Hyponyms:	*idea* | *plan*
	amazon | *ebay*
	journalist | *videographer*
Is-A:	*conspiracy* | *obstruction*
	capability | *repair*
	status | *fame* | *fortune*
Instance Of:	*july* | *month*
	community | *un*
	fedex | *company* | *order*

Table 3. Candidate thesaurus relations.

Synonyms		**Co-Hyponyms**	
administration	government	Amazon	eBay
agency	association	battle	race
imagery	model	flaw	issue
madrassa	school	idea	plan

Is-A		**Instance Of**	
ability	breathing	agency	IAEA
agency	custom	Aisawa	legislator
agreement	deal	Bush	president
cleric	sheik	group	Nirvana

5.2. Similarity Measures. In order to quantify the feasibility of the methodology, we retained only 372 test cases labeled with a specific semantic relation and solved them by Cosine-PMI similarity measure. The results are summarized in Table 4.

Table 4. Accuracy on 372 tests.

Synonyms	**Co-Hyponyms**	**Is-A**	**Instance Of**	**Overall**
75%	**65%**	**58%**	**30%**	**49%**

The contextual similarity measure fails to achieve positive result for the category *Instance Of.* This is no surprise since, in order to be solved, most of the cases in this category reduce to a problem of finding the most salient property associated to a proper name. However on the other categories the method achieves performance comparable to state-of-the-art-systems. This indicates that paraphrases are reliable source of closely related word pairs.

6. Conclusions and Future Work

In this paper, we presented a method for word semantic relation extraction. This approach can be applied to extract different semantic relations, it extracts relations between unfrequent word senses, it limits the search space and it is completely unsupervised.

In particular, as many as 37% of the constructed TOEFL like test cases contain close semantic relations. The methodology is also not hindered by low frequency words and discovered 24 synonymous word pairs not listed in WordNet. Compared to other methods that create long lists of words related in unspecified way, our methodology extracts very short lists of candidates in paradigmatic relation with the head. Those lists can be easily scrutinized by a human expert in computer aided thesaurus construction.

However, the methodology still produces erroneous tests mostly resulting from bad text preprocessing and unreliable part-of-speech tagging. A major improvement can be obtained by the normalization of the corpus i.e. by detecting multiword units and named entities.

As future work, we aim at testing a new alignment technique proposed by [5] who use a combination of local and global biology-based alignment algorithms which deals with sentence reordering.

References

[1] Helena Ahonen-Myka. Finding all frequent maximal sequences in text. In *Proceedings of ICML-99 Workshop on Machine Learning in Text Data Analysis*, pages 11–17, 1999.

[2] Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of ACL*, pages 57–64, 1999.

[3] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, pages 757–766, 2007.

[4] João Cordeiro, Gaël Dias, and Pavel Brazdil. New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software*, 2(4):12–23, 2007.

[5] João Cordeiro, Gaël Dias, and Guillaume Cleuziou. Biology based alignments of paraphrases for sentence compression. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing (ACL-PASCAL / ACL2007)*, pages 177–184, 2007.

[6] James Richard Curran and Marc Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, 2002.

[7] Gaël Dias, Rumen Moraliyski, João Cordeiro, Antoine Doucet, and Helena Ahonen-Myka. Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Journal of Natural Language Engineering, Special Issue on Distributional Lexical Semantics*, 16(04):439–467, 2010.

[8] Zellig Harris. *Mathematical Structures of Language*. Wiley, New York, NY, USA, 1968.

[9] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[10] Kris Heylen, Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation*, pages 3243–3249, 2008.

[11] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting of ACL*, pages 268–275, 1990.

[12] Abraham Kaplan. An experimental study of ambiguity and context. *Mechanical Translation*, 2(2):39–46, 1950.

[13] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[14] Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. Identifying synonyms among distributionally similar words. In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1492–1493, 2003.

[15] Hugo Liu. Montylingua: An end-to-end natural language processor with common sense., 2004. Available at: `web.media.mit.edu/~hugo/montylingua`.

[16] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[17] Egidio Terra and Charlie Clarke. Frequency estimates for statistical word similarity measures. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 165–172, 2003.

Gaël Dias, Rumen Moraliyski, João Cordeiro,
Centre for HLT and Bioinformatics, Department of Computer Science,
University of Beira Interior, 6201-001 - Covilhã, Portugal,
e-mail: ddg@di.ubi.pt, rumen@penhas.di.ubi.pt, jpaulo@di.ubi.pt

Antoine Doucet,
Campus Côte de Nacre, Boulevard du Maréchal Juin
University of Caen, BP 5186 - 14032 - Caen CEDEX, France,
e-mail: doucet@info.unicaen.fr

Helena Ahonen-Myka,
Department of Computer Science,
P.O. Box 68 (Gustaf Hällströmin katu 2b),
FI-00014, University of Helsinki, Finland,
e-mail: helena.ahonen-myka@cs.helsinki.fi