Paraphrase Alignment for Synonym Evidence Discovery

Gintarė Grigonytė
Saarland University
g.grigonyte@hmf.vdu.lt

João Cordeiro, Gaël Dias, Rumen Moraliyski
HULTIG
University of Beira Interior
{jpaulo,ddg,rumen}@di.ubi.pt

Pavel Brazdil
LIAAD/FEP
University of Porto
pbrazdil@liaad.up.pt

Abstract

We describe a new unsupervised approach for synonymy discovery by aligning paraphrases in monolingual domain corpora. For that purpose, we identify phrasal terms that convey most of the concepts within domains and adapt a methodology for the automatic extraction and alignment of paraphrases to identify *paraphrase casts* from which valid synonyms are discovered. Results performed on two different domain corpora show that general synonyms as well as synonymic expressions can be identified with a 67.27% precision.

1 Introduction

Synonymy is a specific type of a semantic relationship. According to (Sowa and Siekmann, 1994), a synonym is a word (or concept) that means the same or nearly the same as another word (or concept). It has been observed that words are similar if their contexts are similar (Freitag et al., 2005) and so synonymy detection has received a lot of attention during the last decades. However, words used in the same context are not necessarily synonyms and can embody different semantic relationships such as hyponyms, meronyms or co-hyponyms (Heylen et al., 2008). In this paper, we introduce a new unsupervised methodology for synonym detection by extracting and aligning paraphrases on normalized domain corpora[1]. In particular, we study a specific structure within aligned paraphrases, *paraphrase casts*, from which valid synonyms are discovered. In fact, we propose a new approach based on the idea that synonyms are substitutable words within a given domain corpus. Results performed on two different domain corpora, the Corpus of Computer Security (COCS) and the Corpus of Cancer Research (COCR), show that general synonyms as well as synonymic expressions can be identified with a 67.27% precision performance.

[1]By normalized, we intend that phrasal terms have been previously identified.

2 Related Work

Automatic synonymy detection has been tackled in a variety of ways which we explain as follows.

2.1 Pattern-based Approaches

This approach to information extraction is based on a technique called *selective concept extraction* as defined by (Riloff, 1993). Selective concept extraction is a form of text skimming that selectively processes relevant text while effectively ignoring surrounding text that is thought to be irrelevant to the domain. The pioneer of pattern-based approaches (Hearst, 1992) has introduced lexico-syntactic patterns to automatically acquire given word semantic relationships. Specific patterns like "*X and other Y*" or "*X such as Y*" were used for hypernym-hyponym detection. Later, the idea was extended and adapted for synonymy by other researchers such as (Roark and Charniak, 1998), (Caraballo, 1999) and (Maynard and Peters, 2009). In general, manual pattern definition is time consuming and requires linguistic skills. Usually, systems based on lexico-syntactic patterns perform with very high precision, but low recall due to the fact that these patterns are rare. However, recent work by (Ohshima and Tanaka,

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 403–411,
Beijing, August 2010

2009) on Web data reported high recall figures. To avoid manual encoding of patterns, many supervised approaches have been proposed as summarized in (Stevenson and Greenwood, 2006).

2.2 Distributional Similarity

Distributional similarity for capturing semantic relatedness is relying on the hypothesis that semantically similar words share similar contexts. These methods vary in the level of supervision from unsupervised to semi-supervised or to supervised. The first type of methods includes the work of (Hindle, 1990), (Lin, 1998) and (Heylen et al., 2008) who used unsupervised methods for detecting word similarities based on shallow-parsed corpora. Others have proposed unsupervised methodologies to solve TOEFL-like tests, instead of discovering synonyms (Turney, 2001), (Terra and Clarke, 2003) and (Freitag et al., 2005). Other researchers, such as (Girju et al., 2004), (Muller et al., 2006), (Wu and Zhou, 2003) and (Wei et al., 2009), have used language or knowledge resources to enhance the representation of the vector space model. Unlike the pattern-based approach, the distributional similarity-based approach shows low precision compared to high recall.

One obvious way to verify all the possible connections between words of the vocabulary employs an exhaustive search. However, comparison based on word usage can only highlight those terms that are highly similar in meaning. This method of representation is usually unable to distinguish between middle strength and weak semantic relations, or detect the relationship between hapax-legomena.

2.3 Hybrid Approaches

More recently, approaches combining patterns and distributional similarity appeared to bring the best of the two methodologies. (Hagiwara et al., 2009) describe experiments that involve training various synonym classifiers. (Giovannetti et al., 2008) use syntactically parsed text and manually composed patterns together with distributional similarity for detecting semantically related words. Finally, (Turney, 2008) proposes a supervised machine learning approach for discovering synonyms, antonyms, analogies and associations. For that purpose, feature vectors are based on frequencies of patterns and classified by a SVM.

2.4 Our Approach

(Van der Plas and Tiedemann, 2006) state that "*People use multiple ways to express the same idea. These alternative ways of conveying the same information in different ways are referred to by the term* paraphrase *and in the case of single words or phrasal terms sharing the same meaning, we speak of synonyms*". Based on this, we propose that in order to discover pairs of semantically related words (in the best case synonyms) that may be used in figurative or rare sense, and as consequence impossible to be identified by the distributional similarity approach, we need to have them highlighted by their **local specific** environment. Here we differ from the pattern-based approach that use **local general** environment. We propose to align paraphrases from domain corpora and discover words that are possibly substitutable for one another in a given context (*paraphrase casts*), and as such are synonyms or near-synonyms. Comparatively to existing approaches, we propose an unsupervised and language-independent methodology which does not depend on linguistic processing[2], nor manual definition of patterns or training sets and leads to higher precision when compared to distributional similarity-based approaches.

3 Normalization of the Corpora

The main goal of our research is to build knowledge resources in different domains that can effectively be used in different NLP applications. As such, precision takes an important part in the overall process of our methodology. For that purpose, we first identify the phrasal terms (or multiword units) present in the corpora. Indeed, it has been shown in many works that phrasal terms convey most of the specific contents of a given domain. Our approach to term extraction is based on linguistic pattern matching and Inverse Document Frequency (IDF) measurements for term

[2]We will see in the next section that we will use linguistic resources to normalize our corpora, but the methodology can be applied to any raw text.

quality assurance as explained in (Avizienis et al., 2009). For that purpose, we present a domain independent hybrid term extraction framework that includes the following steps. First, the text is morphologically annotated with the MPRO system (Maas et al., 2009). Then grammar rules for morphological disambiguation, syntactic parsing and noun phrase detection are applied based on finite-state automata technology, KURD (Carl and Schmidt-Wigger, 1998). Following this, a variant and non-basic term form detection is applied, as well as stop words removal. Then, combining rich morphological and shallow syntactical analysis with pattern matching techniques allows us to extract a wide span of candidate terms which are finally filtered with the well-known IDF measure.

4 Paraphrase Identification

A few unsupervised metrics have been applied to automatic paraphrase identification and extraction (Barzilay and McKeown, 2001) and (Dolan et al., 2004). However, these unsupervised methodologies show a major drawback by extracting quasi-exact or even exact match pairs of sentences as they rely on classical string similarity measures. Such pairs are useless for our research purpose. More recently, (Cordeiro et al., 2007a) proposed the *sumo* metric specially designed for asymmetrical entailed pair identification in corpora which obtained better performance than previously established metrics, even in corpora with exclusively symmetrical entailed paraphrases as in the Microsoft Paraphrase Research Corpus (Dolan et al., 2004). This function states that for a given sentence pair $\langle S_a, S_b \rangle$, having m and n words in each sentence and λ lexical exclusive links (word overlaps, see figure 1) between them, its lexical connection strength is computed as defined in Equations 1 and 2.

$$Sumo(S_a, S_b) = \begin{cases} S(m,n,\lambda) & if \;\; S(m,n,\lambda) < 1 \\ 0 & if \;\; \lambda = 0 \\ e^{-kS(m,n,\lambda)} & otherwise \end{cases} \quad (1)$$

where

$$S(m,n,\lambda) = \alpha \log_2(\tfrac{m}{\lambda}) + \beta \log_2(\tfrac{n}{\lambda})$$
$$\alpha, \beta \in [0,1], \alpha + \beta = 1 \quad (2)$$

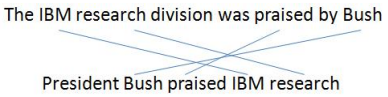

Figure 1: 4 exclusive links between S_a and S_b.

To obtain a paraphrase corpus, we compute all sentence pairs similarities $Sumo(S_a, S_b)$ and select only those pairs exceeding a given threshold, in our case $threshold = 0.85$, which is quite restrictive, ensuring the selection of pairs strongly connected[3].

However, to take into account the normalization of the corpus, little adjustments had to be integrated in the methodology proposed in (Cordeiro et al., 2007a). Indeed, the original $Sumo(.,.)$ function was under-weighting links that occurred between phrasal terms such as "*molecular laboratory*" or "*renal cancer*". So, instead of counting the number of lexical links among sentences, each link weights differently according to the word length in the connection, hence connections of longer words will result in a larger value. For example, in figure 1, instead of having $\lambda = 4$, we count $\lambda = 3 + 8 + 7 + 4 = 22$. This adjustment is important as multi-word units are treated as longer words in the corpus. This modification has also, as a side effect, under-evaluation of functional words which usually follow the Zipf's Law and give more importance to meaningful words in the paraphrase extraction process.

5 Paraphrase Alignment

In order to usefully explore the evidence synonymy from paraphrases, sentence alignment techniques must be applied to paraphrases in order to identify *paraphrase casts*, i.e., substitutable word pairs as shown in figure 2. As we can see, the paraphrase cast includes three parts: the left segment (context), a middle segment and the right segment (context). In our figure the left and right segments (contexts) are identical.

In the context of DNA sequence alignment, two main algorithms have been proposed: (1) the Needleman-Wunsch algorithm (Needleman and

[3]Further details about the sumo metric are available in (Cordeiro et al., 2007a).

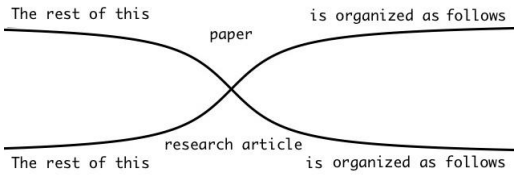

Figure 2: A paraphrase cast.

Wunsch, 1970) based on dynamic programming which outputs a unique global alignment and (2) the Smith-Waterman (SW) algorithm (Smith and Waterman, 1981) which is an adaptation of the previous algorithm and outputs local alignments. In the context of NLP, (Cordeiro et al., 2007a) proposed a combination of both algorithms depending on the structure of paraphrase. However, since any local alignment is a candidate for *paraphrase casts*, the SW algorithm revealed itself more appropriate and was always chosen. The SW alignment algorithm uses dynamic programming to compute the optimal local alignments between two sequences[4]. This process requires first the definition of an alignment matrix (function), which governs the likelihood of alignment of two symbols. Thus we first build a matrix H such that $H(i,0) = 0$ and $H(0,j) = 0$, for $0 \leq i \leq m$, and $0 \leq j \leq n$, where m and n are the number of words in the paraphrase sentences. The rest of the H elements are recursively calculated as in Equation 3 where $gs(.,.)$ is the gap-scoring function and S_{a_i} (resp. S_{b_j}) represents the i^{th} (resp. j^{th}) word of sentence S_a (resp. S_b).

$$H(i,j) = max \begin{cases} 0 & \\ H(i-1,j-1) + gs(S_{a_i}, S_{b_j}) & \text{MMisatch} \\ H(i-1,j) + gs(S_{a_i}, \_) & \text{Deletion} \\ H(i,j-1) + gs(\_, S_{b_j}) & \text{Insertion} \end{cases} \quad (3)$$

Obviously, this algorithm is based on an alignment function which exploits the alignment likelihood between two alphabet symbols. For DNA sequence alignments, this function is defined as a mutation matrix, scoring gene mutation and gap alignments. In our case, we define the gap-scoring function $gs(.,.)$ in Equations 4 and 5 which prioritize the alignment of specific domain key terms i.e., single match, or key expressions i.e., phrasal match, (reward 50), as well as lexically similar[5] words such as "*programme*" and "*programming*" for example. In particular, for these similar words an adaptation of the well known *Edit Distance* is used, the $c(.,.)$ function (5), which is explained in more detail in (Cordeiro et al., 2007b).

$$gs(S_{a_i}, S_{b_j}) = \begin{cases} -1 & \text{if } (S_{a_i} = -) \text{ and } (S_{b_j} \neq -) \\ -1 & \text{if } (S_{b_j} = -) \text{ and } (S_{a_i} \neq -) \\ 10 & \text{Single Match} \\ 50 & \text{Phrasal Match} \\ c(S_{a_i}, S_{b_j}) & \text{Mismatch} \end{cases} \quad (4)$$

where

$$c(S_{a_i}, S_{b_j}) = -\frac{edist(S_{a_i}, S_{b_j})}{\epsilon + maxseq(S_{a_i}, S_{b_j})} \quad (5)$$

To obtain local alignments, the SW algorithm is applied, using the alignment function defined with $H(.,.)$ in 3. The alignment of the paraphrase in figure 2 would give the result in figure 3.

```
The rest of this paper ---------------- is organized as follows
The rest of the  ----- research article is organized as follows
```

Figure 3: An alignment.

6 Paraphrase Casts

In order to discover synonyms, we are looking for special patterns from the aligned paraphrase sentences, which naturally give us more evidence for the existence of equivalent terms or expressions. Due to the topological aspect of such patterns, we decided to name them *paraphrase casts*, or just *casts* as shown in figure 2. As we have mentioned earlier, the paraphrase cast includes three parts: the left segment (*contextL*), a middle segment and the right segment (*contextR*). In the following example the left and right segments (contexts) are identical, but the middle segment includes **different** misaligned sequences of words, represented by *wordSa* and *wordSb*.

```
contextL wordSa ----- contextR
contextL ----- wordSb contextR
```

[4]In our case, the two sequences are the two sentences of a paraphrase

[5]This is why we have in equation 3 the label "Mismatch", where "mismatch" means different yet lexically near words.

We can attribute different levels of confidence to different paraphrase casts. Indeed, the larger the contexts and the smaller the misaligned sequences are, the more likely it is for single or phrasal terms to be synonyms or near-synonyms. Note that in the cast shown in figure 3, each context has a significant size, with four words on each side, and the misaligned segments are in fact equivalent expressions i.e. "*paper*" is a synonym of "*research article*". In the analyzed domain these expressions are equivalent and interchangeable and appear to be interchangeable in other domains. For the purpose of this paper, we only take into account the casts where the misaligned sequences of words contain only one word or one multi-word unit in each sentence. That is, we have a one-to-one match. However, no constraints are imposed on the contexts[6]. So, all casts are computed and analyzed for synonym discovery and results are provided in the next section.

7 Experiments

To evaluate our methodology we have used two different corpora, both from scientific domains built from abstracts of publications (see Table 1). The corpus of computer security (COCS) is a collection of 4854 abstracts on computer security extracted from the IEEE (http://ieee.rkbexplorer.com/) repository[7]. The corpus of cancer research (COCR) contains 3334 domain specific abstracts of scientific publications extracted from the PubMed[8] on three types of cancer: (1) the mammary carcinoma register (COCR1) consisting of 1500 abstracts, (2) the nephroblastoma register (COCR2) consisting of 1500 abstracts, and (3) the rhabdoid tumor register (COCR3) consisting of 334 abstracts. From the paraphrase casts, we were able to automatically extract, without further processing, single synonymous word pairs, as well as synonymic multi-word units, as can be seen in Table 2. For that purpose we have used specific paraphrase casts, whose aim was to privilege precision to

[6]This issue will be discussed in the next section.

[7]An example of an abstract can be viewed at: http://ieee.rkbexplorer.com/description/publication-00534618

[8]http://www.ncbi.nlm.nih.gov/pubmed

Corpus	COCS	COCR1	COCR2	COCR3
Tokens	412.265	336.745	227.477	46.215
Sentences	18.974	15.195	10.575	2.321
Aligned Pairs	589	994	511	125
Casts without filter	320	10.217	2.520	48
Casts with filter	202	361	292	16

Table 1: Corpora

recall. In particular, we have removed all casts which contained numbers or special characters i.e. casts with filter in Table 1. However, no constraints were imposed on the frequency of the casts. Indeed, all casts were included even if their overall frequency was just one. Although

Synonym (COCS)	**Complementary**
frequency tuning	frequency control
attack consequences	attack impact
error-free operation	error free operation
pseudo code	pseudo algorithm
tolerance	resilience
package loss	message loss
adjustable algorithm	context-aware algorithm
helpful comment	valuable comment
Synonym (COCR)	**Complementary**
childhood renal tumor	childhood kidney tumor
hypertrophy	growth
doxorubicin	vincristine
carcinomas of the kidney	sarcoma of the kidney
metastasis	neoplasm
renal tumor	renal malignancy
neoplastic thrombus	tumor thrombus
vincristine	adriamycin

Table 2: Synonyms for COCS

most of the word relationships were concerned with synonymy, the other ones were not just errors, but lexically related words, namely examples of antonymy, hyperonymy/hyponymy and associations as shown in Table 3. In the evaluation, we

Antonym	**Complementary**
positive sentinel nodes	negative sentinel nodes
higher bits	lower bits
older version	newer version
Hypernym	**Hyponym**
Multi-Tasking Virtual Machine	Java Virtual Machine
therapy	chemotherapy
hormone breast cancer	estrogen breast cancer
Association	**Complementary**
performance	reliability
statistical difference	significant difference
relationship	correlation

Table 3: Other Word Semantic Relationships.

have focused on the precision of the method. The evaluation of the extracted pairs was performed manually by two domain experts. In fact, four

different evaluations were carried out depending on whether the adapted $S(.,.)$ measure was used (or not) and whether the normalization of the corpora was used (or not). The best results were obtained in all cases for the adapted $S(.,.)$ measure with the multi-word units. Table 4 shows also the worst results for the COCS as a baseline (COCS (1)), i.e. non-adapted $S(.,.)$ and non-normalized corpus. For the rest of the experiments we always present the results with the adapted $S(.,.)$ measure and normalized corpus.

Corpus	**COCS (1)**	**COCS (2)**	
Precision	54.62%	71.29%	
Extracted Synonyms	130	144	
Errors	108	58	
Corpus	**COCR1**	**COCR2**	**COCR3**
Precision	69.80%	61.30%	75.00%
Extracted Synonyms	252	178	12
Errors	109	111	4

Table 4: Overall Results

7.1 Discussion

Many results have been published in the literature, especially tackling the TOEFL synonym detection problem which aims at identifying the correct synonym among a small set of alternatives (usually four). For that purpose, the best precision rate has been reached by (Turney et al., 2003) with 97.50% who have exploited many resources, both statistical and linguistic. However, our methodology tackles **a different problem**. Indeed, our goal is to automatically extract synonyms from texts. The published works toward this direction have not reached so good results. One of the latest studies was conducted by (Heylen et al., 2008) who used distributional similarity measures to extract synonyms from shallow parsed corpora with the help of unsupervised methods. They report that "*the dependency-based model finds a tightly related neighbor for 50% of the target words and a true synonym for 14%*". So, it means that by comparing all words in a corpus with all other words, one can expect to find a correct semantic relationship in 50% of the cases and a correct synonym in just 14%. In that perspective, our approach reaches higher results. Moreover, (Heylen et al., 2008) use a frequency cut-off which only selects features that occur at least five times together with the target word. In our case, no frequency threshold is imposed to enable extraction of synonyms with low frequency, such as *hapax legomena*. This situation is illustrated in figure 4. We note that most of the candidate pairs appear only once in the corpus.

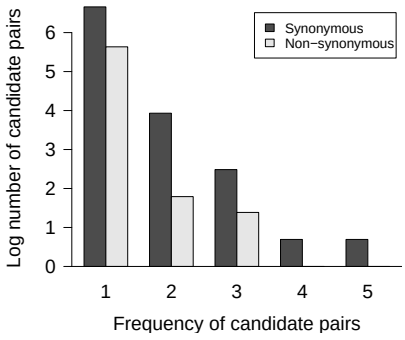

Figure 4: Synonyms Frequency Distribution.

In order to assess the quality of our results, we calculated the similarity between all extracted pairs of synonyms following the distributional analysis paradigm as in (Moraliyski and Dias, 2007) who build context[9] feature vectors for noun synonyms. In particular, we used the cosine similarity measure and the Loglike association measure (Dunning, 1993) as the weighting scheme of the context features. The distribution of the similarity measure for all noun synonyms (62 pairs) is shown in figure 5.

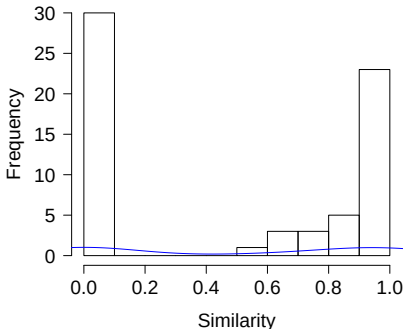

Figure 5: Synonym Pairs Similarity Distribution.

The results clearly show that all extracted synonyms are highly correlated in terms of context.

[9]In this case, the contexts are the surrounding nouns, verbs and adjectives in the closest chunks of a shallow parsed corpus.

Nearly half of the cases have similarities higher than 0.5. It is important to notice that a specific corpus[10] was built to calculate as sharply as possible the similarity measures as it is done in (Moraliyski and Dias, 2007). Indeed, based on the COCS and the COCR most statistics were insignificant leading to zero-valued features. This situation is well-known as it is one of the major drawbacks of the distributional analysis approach which needs huge quantities of texts to make secure decisions. So we note that applying the distributional analysis approach to such small corpora would have led to rather poor results. Even with an adapted corpus, figure 5 (left-most bar) shows that there are no sufficient statistics for 30 pairs of synonyms. Although the quality of the extracted pairs of synonyms is high, the precision remains relatively low with 67.27% precision on average. As we pointed out in the previous section, we did not make any restrictions to the left and right contexts of the casts. However, the longer these contexts are, compared to the misaligned sequence of words, the higher the chance is that we find a correct synonym. Table 5 shows the average lengths of both cast contexts for synonyms and erroneous pairings, in terms of words (WCL) and characters (CCL). We also provide the ratio (R) between the character lengths of the middle segment (i.e. misaligned character sequences) and the character lengths of the cast contexts (i.e. right and left sizes of equally aligned character sequences). It is

Type	WCL	CCL	R
Synonyms	2.70	11.67	0.70
Errors	2.45	8.05	0.55

Table 5: Average Casts Contexts Lengths

clear that a more thorough study of the effects of the left and right contexts should be carried out to improve precision of our approach, although this may be to the detriment of recall. Based on the results of the ratio R[11], we note that the larger the cast context is compared to the cast content, the more likely it is that the misaligned words are synonyms.

[10]This corpus contains 125.888.439 words.

[11]These results are statistically relevant with $p - value < 0.001$ using the Wilcoxon Rank-Sum Test.

8 Conclusions

In this paper we have introduced a new unsupervised methodology for synonym detection that involves extracting and aligning paraphrases on normalized domain corpora. In particular, we have studied a specific structure within aligned paraphrases, *paraphrase casts*, from which valid synonyms were discovered. The overall precision was 71.29% for the computer security domain and 66.06% for the cancer research domain. This approach proved to be promising for extracting synonymous words and synonymic multiword units. Its strength is the ability to effectively work with small domain corpora, without supervised training, nor definition of specific language-dependent patterns. Moreover, it is capable to extract synonymous pairs with figurative or rare senses which would be impossible to identify using the distributional similarity approach. Finally, our approach is completely language-independent as it can be applied to any raw text, not obligatorily normalized corpora, although the results for domain terminology may be improved by the identification of phrasal terms.

However, further improvements of the method should be considered. A measure of quality of the *paraphrase casts* is necessary to provide a measure of confidence to the kind of extracted word semantic relationship. Indeed, the larger the contexts and the smaller the misaligned sequences are, the more likely it is for single or phrasal terms to be synonyms or near-synonyms. Further work should also be carried out to differentiate the acquired types of semantically related pairs. As it is shown in Table 3, some of the extracted pairs were not synonymic, but lexically related words such as antonyms, hypernyms/hyponyms and associations. A natural follow-up solution for discriminating between semantic types of extracted pairs could involve context-based classification of acquired *casts* pairs. In particular, (Turney, 2008) tackled the problem of classifying different lexical information such as synonymy, antonymy, hypernymy and association by using context words. In order to propose a completely unsupervised methodology, we could also follow the idea of (Dias et al., 2010) to automatically construct small

TOEFL-like tests based on sets of *casts* which could be handled with the help of different distributional similarities.

Acknowledgments

We thank anonymous reviewers whose comments helped to improve this paper. We also thank IFOMIS institute (Saarbrucken) and the ReSIST project for allowing us to use the COCR and COCS corpora.

References

Avizienis, A., Grigonyte, G., Haller, J., von Henke, F., Liebig, T., and Noppens, O. 2009. *Organizing Knowledge as an Ontology of the Domain of Resilient Computing by Means of NLP - An Experience Report.* In Proc. of the 22th Int. FLAIRS Conf. AAAI Press, pp. 474-479.

Barzilay, R. and McKeown, K. R. 2001. *Extracting Paraphrases from a Parallel Corpus*. In Proc. of the 39th meeting of ACL, pp. 50-57.

Caraballo, S. A. 1999. *Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text*. In Proc. of 37th meeting of ACL 1999, pp 120-126.

Carl, M., and Schmidt-Wigger, A. 1998. *Shallow Post Morphological Processing with KURD*. In Proc. of the Conf. on New Methods in Language Processing.

Cordeiro, J.P., Dias, G. and Brazdil, P. 2007. *Learning Paraphrases from WNS Corpora.* In Proc. of the 20th Int. FLAIRS Conf. AAAI Press, pp. 193-198.

Cordeiro, J.P., Dias, G. and Cleuziou, G. 2007. *Biology Based Alignments of Paraphrases for Sentence Compression.* In Proc. of the 20th meeting of ACL, workshop PASCAL, pp. 177-184.

Dias, G., Moraliyski, R., Cordeiro, J.P., Doucet, A. and Ahonen-Myka, H. 2010. *Automatic Discovery of Word Semantic Relations using Paraphrase Alignment and Distributional Lexical Semantics Analysis*. In Journal of Natural Language Engineering, Cambridge University Press. ISSN 1351-3249, pp. 1-26.

Dolan, B., Quirk, C. and Brockett, C. 2004. *Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources*. In Proc. of the 20th int. Conf. on Computational Linguistics.

Dunning T. D 1993. *Accurate Methods for the Statistics of Surprise and Coincidence*. In Computational Linguistics, 19(1), pp. 61-74.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R. and Wang, Z. 2005. *New Experiments in Distributional Representations of Synonymy*. In Proc. of 9th conf. on Computational Natural Language Learning, pp. 25-32.

Giovannetti, E., Marchi, S. and Montemagni, S. 2008. *Combining Statistical Techniques and Lexico-Syntactic Patterns for Semantic Relations Extraction from Text*. In Proc. of the 5th Workshop on Semantic Web Applications and Perspectives.

Girju, R., Giuglea, A. M., Olteanu, M., Fortu, O., Bolohan, O. and Moldovan, D. 2004. *Support Vector Machines Applied to the Classification of Semantic Relations in Nominalized Noun Phrases*. In Proc. of the HLT-NAACL Workshop on Computational Lexical Semantics, pp. 68-75.

Hagiwara, M. O. Y. and Katsuhiko, T. 2009. *Supervised Synonym Acquisition using Distributional Features and Syntactic Patterns*. In Information and Media Technologies 4(2), pp. 558-582.

Hearst, M. A. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proc. of the 14th conf. on Computational Linguistics, pp. 539-545.

Heylen, K., Peirsman, Y., Geeraerts, D. and Speelman, D. 2008. *Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms*. In Proc. of the 6th LREC.

Hindle, D. 1990. *Noun Classification from Predicate-Argument Structures*. In Proc. of the 28th meeting of ACL, pp. 268-275.

Inkpen, D. 2007. *A Statistical Model for Near-Synonym choice*. In ACM Trans. Speech Lang. Process. 4(1), 1-17.

Kiefer, J. 1953. *Sequential Minimax Search for a Maximum*. In Proc. of the American Mathematical Society 4, pp. 502-506.

Lin, D. 1998. *Automatic Retrieval and Clustering of Similar Words*. In Proc. of the 17th Int. Conf. on Computational Linguistics, pp. 768-774.

Maas, D., Rosener, Ch., and Theofilidis, A. 2009. *Morphosyntactic and Semantic Analysis of Text: The MPRO Tagging Procedure*. Workshop on systems and frameworks for computational morphology. Springer., pp. 76-87.

Maynard, D. F. A. and Peters, W. 2009. *Using lexico-syntactic Ontology Design Patterns for Ontology Creation and Population*. In Proc. of the Workshop on Ontology Patterns.

Moraliyski, R., and Dias, G. 2007. *One Sense per Discourse for Synonym Detection*. In Proc. of the Int. Conf. On Recent Advances in NLP, Bulgaria, pp. 383-387.

Muller, P., Hathout, N. and Gaume, B. 2006. *Synonym Extraction Using a Semantic Distance on a Dictionary*. In Proc. of the 1st Workshop on Graph-Based Methods for NLP, pp. 65-72.

Needleman, S. B. and Wunsch, C. D. 1970. *A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins*. In Journal of Molecular Biology 48(3), pp. 443-453.

Ohshima, H. and Tanaka, K. 2009. *Real Time Extraction of Related Terms by Bi-directional lexico-syntactic Patterns from the Web*. In Proc. of the 3rd Int. Conf. on Ubiquitous Information Management and Communication, pp. 441-449.

Riloff, E. 1993. *Automatically Constructing a Dictionary for Information Extraction Tasks*. In Proc. of the 11th Nat. Conf. on Artificial Intelligence, pp. 811-816.

Roark, B. and Charniak, E. 1998. *Noun-phrase Co-occurrence Statistics for Semiautomatic Semantic Lexicon Construction*. In Proc. of the 17th Int. Conf. on Computational Linguistics, pp. 1110-1116.

Smith, T. and Waterman, M. 1981. *Identification of Common Molecular Subsequences*. In Journal of Molecular Biology 147, pp. 195-197.

Sowa, J. F. and Siekmann, J. H. 1994. *Conceptual Structures: Current Practices*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Stevenson, M. and Greenwood, M. 2006. *Comparing Information Extraction Pattern Models*. In Proc. of the Workshop on Information Extraction Beyond the Document, ACL, pp. 29-35.

Terra, E. and Clarke, C. 2003. *Frequency Estimates for Statistical Word Similarity Measures*. In Proc. of HTL/NAACL 2003, pp. 165-172.

Turney, P. D. 2001. *Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.* Lecture Notes in Computer Science, 2167, pp. 491-502.

Turney, P. D., Littman, M. L., Bigham, J. and Shnayder, V. 2003. *Combining Independent Modules in Lexical Multiple-choice Problems*. In Recent Advances in NLP III: Selected Papers, pp. 101-110.

Turney, P. D. 2008. *A Uniform Approach to Analogies, Synonyms, Antonyms and Associations*. In Proc .of the 22nd Int. Conf. on Computational Linguistics, pp. 905-912.

Van der Plas, L. and Tiedemann, J. 2006. *Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity*. In Proc. of the 21st Int. Conf. on Computational Linguistics, pp. 866-873.

Wei, X., Peng, F., Tseng, H., Lu, Y. and Dumoulin, B. 2009. *Context Sensitive Synonym Discovery for Web Search Queries*. In Proc. of the 18th ACM conference on Information and Knowledge Management, pp. 1585-1588.

Wu, H. and Zhou, M. 2003. *Optimizing Synonym Extraction Using Monolingual and Bilingual Resources*. In Proc. of the 2nd Int. Workshop on Paraphrasing, pp. 72-79.