Semantic Clustering of Pivot Paraphrases

Marianna Apidianaki[1], Emilia Verzeni[2], Diana McCarthy[3]

1. LIMSI-CNRS, Rue John von Neumann, 91403 Orsay, France
2. University Paris Diderot, 5 rue Thomas-Mann, 75205 Paris cedex 13, France
3. Department of Theoretical and Applied Linguistics (DTAL), University of Cambridge, UK
marianna@limsi.fr, emilia.ve@gmail.com, diana@dianamccarthy.co.uk

Abstract

Paraphrases extracted from parallel corpora by the pivot method (Bannard and Callison-Burch, 2005) constitute a valuable resource for multilingual NLP applications. In this study, we analyse the semantics of unigram pivot paraphrases and use a graph-based sense induction approach to unveil hidden sense distinctions in the paraphrase sets. The comparison of the acquired senses to gold data from the Lexical Substitution shared task (McCarthy and Navigli, 2007) demonstrates that sense distinctions exist in the paraphrase sets and highlights the need for a disambiguation step in applications using this resource.

Keywords: pivot paraphrasing, sense clustering, parallel corpora

1. Introduction

The *pivot* method offers an inexpensive way to acquire semantic information in languages with parallel corpora (Bannard and Callison-Burch, 2005). The main assumption underlying this method is that two phrases that are paraphrases of each other may be translated in the same way in a foreign language (Dyvik, 1998). This translation information, available in the phrase table of state-of-the-art Statistical Machine Translation systems, makes the automatic acquisition of paraphrase sets straightforward. Pivot paraphrases are thus widely used in Machine Translation (MT): they serve to translate out-of-vocabulary (OOV) words (Callison-Burch et al., 2006) and can help to establish sense correspondences during MT evaluation (Zhou et al., 2006; Madnani et al., 2007; Snover et al., 2010; Denkowski and Lavie, 2010).

Nevertheless, the paraphrases induced for an ambiguous word might carry the word's different meanings, similar to its translations. Our aim is to investigate the semantics of unigram paraphrases identified by the pivot method and to explore the extent to which they are semantically equivalent or whether there is evidence of ambiguity necessitating an additional processing stage in applications exploiting such data. We focus on paraphrases acquired for English words through French and cluster them by meaning using a graph-based approach to sense induction. The acquired senses are compared to a publicly available resource, the Lexical Substitution gold data (McCarthy and Navigli, 2007) (hereafter LexSub), leveraged using the Cluster Projection (CP) algorithm by Bansal et al. (2012). Our assumption is that a high similarity of our clustering to the gold sense groupings would denote the existence of sense distinctions in the paraphrase sets and would highlight the need for a clustering step in applications exploiting this resource. We further explain the motivation behind this work in the next Section. In Section 3, we describe the experimental setup used for clustering. The Cluster Projection method and the evaluation results are presented in Section 4, before concluding.

2. Discovering senses in paraphrase sets

2.1. The pivot method

The pivot method is a widely used technique for extracting paraphrases from bilingual parallel corpora (Bannard and Callison-Burch, 2005). Previous work extracted paraphrases from monolingual parallel corpora by identifying divergent strings in identical surrounding contexts occuring in aligned monolingual sentences (Barzilay and McKeown, 2001; Barzilay and Lee, 2003). The pivot method uses instead phrases in the foreign language of a bilingual parallel corpus as pivots to identify paraphrases in the source language. The method exploits information in the translation table of phrase-based Statistical Machine Translation systems (Koehn et al., 2003). Source language phrases are considered to be potential paraphrases of each other if they share translations in the other language. Using this technique, multiple candidate paraphrases can be extracted for each source phrase.

Each of the extracted candidate paraphrases for a source phrase is assigned a score defined by the translation model probabilities between source and target language phrases and used for ranking. However, since a source phrase can be translated by multiple foreign phrases in the parallel corpus, the paraphrase score is calculated by summing over the different target language phrases. The source phrases that are aligned with these different foreign phrases, which might indicate different senses, are thus mingled. Therefore paraphrases that reflect different senses of the original phrase are included in the same candidate paraphrase set.

Later work by Callison-Burch (2008) proposes a refined paraphrasing technique which decreases the noise present in the generated paraphrase sets. He proposes to use an additional constraint during paraphrase extraction which ensures that the obtained paraphrases are of the same syntactic type as the phrase that they are paraphrasing. Nevertheless, the paraphrase sets might still group together paraphrases corresponding to different senses.

2.2. Ambiguity in paraphrase sets

We focus our study on unigram paraphrases extracted by the variant of the pivot method that uses syntactic constraints (Callison-Burch, 2008) which ensure that the paraphrases pertain to the same grammatical category as the target word. Here is the paraphrase set (P) acquired for the noun *figure*: {*number, amount, chapter, one, personality, statistic, percentage, person, sum*}. After eliminating the noise present in this paraphrase set (*chapter, one*), we can establish a clear semantic distinction between the paraphrases that carry the "numerical figure" and the "person" sense. This difference in meaning could pose problems to matching and substitution in context. It would, for instance, be erroneous to consider *personality* as a substitute of *figure* in the following sentence:

> The *figure* in the first reading, which the House voted yesterday, was EUR 5.5 billion.

and it could lead to mistaken judgments of translation equivalence in an MT setting.

This ambiguity might be resolved in longer n-grams where information supplied by the context can guide the disambiguation. For instance, the "person" sense of *figure* does not appear in the paraphrase set of *final figure*: {*final sum, end result, final version, final result, definitive solution, total amount, grand total, final resolution, statistics, number*}. In this study, we focus on unigram pivot paraphrases (synonyms) where ambiguities are more prevalent and explore ways of unveiling hidden sense distinctions. Following Di Marco and Navigli (2013), we adopt a graph-based sense induction approach.

3. Experimental setup

3.1. Data

To analyse the semantics of pivot paraphrases we exploit distributional information from the parallel corpus that served to extract the paraphrases, i.e. the English-French part of Europarl (Koehn, 2005).[1] The corpus is lemmatised and tagged by part-of-speech on both sides using the TreeTagger (Schmid, 1994). We focus on the target words of the LexSub test data (McCarthy and Navigli, 2007) which consists of 50 nouns, 44 verbs, 47 adjectives and 30 adverbs. The LexSub gold data used for evaluation provides alternative paraphrases of these target words in sentential context. A set of paraphrases with frequency of response from the annotators is provided for each target word in each context as shown in Figure 1, which contains the gold annotations for different instances of the noun *figure*. Each set of substitutes describes the sense of the corresponding target word instance. For example, the gold annotation for instance #1846 of *figure*:

> Some suggest that children are simply more willing to accept the values of parents and teachers when these authority *figures* are affectionate.

[1]The pivot paraphrases are available here: `http://cs.jhu.edu/~ccb/howto-extract-paraphrases.html`

#1841 ::	entity 1;organisation 1;character 1;representative 1; individual 1;
#1842 ::	number 4;statistic 1;numeral 1;
#1843 ::	number 3;amount 2;statistic 1;
#1844 ::	diagram 2;picture 2;illustration 1;people 1;image 1;
#1845 ::	number 3;statistic 1;account 1;value 1;
#1846 ::	leader 1;person 1;character 1;person of authority 1; representative 1;individual 1;
#1847 ::	number 3;statistic 1;amount 1;calculation 1;
#1848 ::	character 3;person 1;personage 1;image 1;
#1849 ::	number 3;statistic 2;data 1;total 1;
#1850 ::	diagram 4;illustration 1;picture 1;

Figure 1: Gold LexSub annotations for noun *figure*.

contains the following paraphrases: {leader 1;person 1;character 1;person of authority 1; representative 1;individual 1}. Each paraphrase has been proposed once by the annotators.[2]

From the LexSub data used for evaluation, we obtain a listing of 'synonym' sets for each target word which may overlap but are not subsets of one another and describe different and possibly related senses.

3.2. Reference clusters

The LexSub annotations might contain paraphrases that were not found in the parallel corpus and which were not extracted by the pivot method. For instance, most of the annotations available for instance #1847 of *figure* were identified as pivot paraphrases (*number, amount, statistic*) but *calculation* was not.

Given the paraphrase set P produced for a target word (w) by the pivot method, we filter P to keep only those paraphrases that exist in the LexSub data. This filtering facilitates the comparison to the gold standard as it keeps only information from LexSub that is relevant for evaluation (i.e. words that appear in our clusters).

To create reference clusters appropriate for evaluating the clustering output, we leverage the LexSub dataset by the Cluster Projection (CP) method of Bansal et al. (2012). The CP algorithm takes as input the paraphrase set P of some target word w. Each paraphrase $p \in P$ belongs to some of the LexSub synonym sets D_t where each set $C \in D_t$ contains paraphrases that may or may not be paraphrases of w. The CP algorithm constructs a source-specific set B for each synonym set C, which contains only paraphrases of w. We apply the CP algorithm to create a set of reference synsets. Furthermore, we exclude from the paraphrase sets the ten most frequent content words in the BNC (*be, have, do, not, say, go, get, make, out, up*) which are highly polysemous.

3.3. Vector creation

Each paraphrase $p \in P$ of a word w is represented as a feature vector describing its distributional context in the source (English) side of the corpus. The retained features

[2]We do not use the annotator frequency information alongside the paraphrases, but treat each paraphrase as equally legitimate for the context.

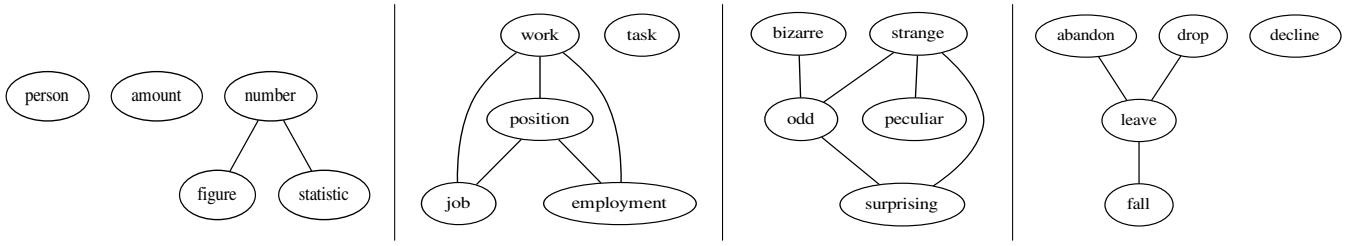

Figure 2: Coarse-grained senses induced for *figure* and *job* (nouns), *strange* (adjective) and *drop* (verb).

(f's) are content words (nouns, verbs, adjectives, adverbs) found in the same sentence as p in Europarl.

The context feature vector of a paraphrase p is constructed by aggregating the frequency counts of each feature f in the contexts of p. We wish to assign higher weights to features that appear less frequently over the entire set of paraphrases, the idea being that a feature with a high surprise (rare) shared by two vectors is more significant regarding their similarity than a more frequent feature. We use Positive Pointwise Mutual Information (PPMI), which is calculated like Pointwise Mutual Information (PMI) (Church and Hanks, 1991) but by replacing negative values with 0 (Turney and Pantel, 2010). The PMI scores are computed from the previously gathered frequency counts as follows:

$$PMI(p,f) = \log \frac{P(p,f)}{P(p)P(f)} \quad (1)$$

where p is a paraphrase, f is a feature retained from its context (i.e. a neighbouring word), $P(p)$ and $P(f)$ are their respective estimated probability in the corpus and $P(p,f)$ is the estimated probability of their co-occurrence. Given that PMI introduces a bias towards infrequent features, we apply the smoothing procedure proposed by Pantel and Lin (2002) which consists of multiplying each PMI score with a discounting factor. This weighting reduces PMI values of rare events relative to frequent ones.

We calculate the similarity of two paraphrases in a paraphrase set P as the cosine of the angle between the corresponding vectors. Let x and y be two paraphrase vectors, $x = < x_1, ..., x_i >$ and $y = < y_1, ..., y_i >$, their cosine is

$$cos(x,y) = \frac{x}{||x||} \cdot \frac{y}{||y||} \quad (2)$$

We calculate the cosine distance between each pair of vectors as: $1 - cos(x,y)$.

3.4. Graph-based sense induction

For each paraphrase set P, we create a graph whose nodes are the paraphrases in P and two paraphrases are linked if they satisfy the following criteria:

1. **alignment**: the two English paraphrases share at least one French translation in the parallel corpus;

2. **similarity**: the similarity score of the paraphrases exceeds a threshold dynamically defined for each target word.

A constraint similar to the alignment criterion above was used in an experiment by Bannard and Callison-Burch (2005) aimed at controlling for word sense (i.e. limit the candidate paraphrases to the same sense as the original phrase). Contrary to the standard method which calculates paraphrase scores by summing over different target language phrases, in this experiment the candidate paraphrases were restricted to those that aligned with the same target language phrase. The accuracy of the extracted paraphrases increased dramatically when word sense was controlled in this way. Nevertheless, paraphrases aligned to the same ambiguous target phrase may still carry different senses. Furthermore, filters requiring alignment to the same target language phrase will reduce the coverage of the resource. The 'disambiguated' version of the resource is not used in subsequent experiments (Callison-Burch, 2008), presumably because of this issue of low coverage. We wish in future to explore approaches to partitioning paraphrase sets which are sensitive to sense but retain as much paraphrase data as possible given the evidence from multilingual, not just bilingual, data.

The similarity threshold is calculated locally for each target word (w) using an iterative procedure (Apidianaki and He, 2010). The threshold (T) is initially set to the mean of the cosine distance of w's paraphrase pairs. The paraphrase pairs are divided in two sets ($G1$, $G2$) depending on whether their distance exceeds or is inferior to the threshold T. The average of the distance scores of the paraphrase pairs in each set is computed ($m1$ and $m2$) and a new threshold is calculated that is the average of $m1$ and $m2$ (T' = (m1 + m2)/2). The new threshold serves to separate again the paraphrase pairs into two sets, a new threshold is calculated and the procedure is repeated until convergence.

The graph obtained for a paraphrase set is then partitioned into cliques and high density components which correspond to senses of different granularity. In Figure 2, we describe the coarse-grained senses (components) identified for some target words in our lexical sample. Distinct senses are described by disjoint components, as in the case of *figure*, *job* and the verb *drop*, while for the adjective *strange* one core sense is found.

4. Evaluation

4.1. Metrics

We evaluate our clustering using standard metrics from the 2010 SemEval WSI task (Manandhar et al., 2010): the V-measure (Rosenberg and Hirschberg, 2007) and the paired

Metric	Data	Results	#cl	1c1par	#cl	1c1word	#cl	Random (5))	#cl
F-score	**Cliques**	0.545	2.06	0.466	2.57	0.681	1	0.534	1.76
	Components	0.629	1.71	0.466	2.57	0.681	1	0.534	1.76
V-measure	**Cliques**	0.582	2.06	0.583	2.57	0.526	1	0.545	1.76
	Components	0.601	1.71	0.583	2.57	0.526	1	0.545	1.76

Table 1: Evaluation results.

F-score (Artiles et al., 2009).[3] V-Measure assesses the quality of a clustering by measuring its *homogeneity* (h) and its *completeness* (c). Homogeneity refers to the degree that each cluster consists of data points primarily belonging to a single gold standard class, while completeness refers to the degree that each gold standard class consists of data points primarily assigned to a single cluster. V-Measure is the harmonic mean of h and c.

$$VM = \frac{2 \cdot h \cdot c}{h + c} \tag{3}$$

In the paired F-Score (Artiles et al., 2009) evaluation, the clustering problem is transformed into a classification problem (Manandhar et al., 2010). A set of instance pairs is generated from the automatically induced clusters, which comprises pairs of the instances found in each cluster. Similarly, a set of instance pairs is created from the gold standard classes, containing pairs of the instances found in each class. *Precision* is then defined as the number of common instance pairs between the two sets to the total number of pairs in the clustering solution (cf. formula 4). *Recall* is defined as the number of common instance pairs between the two sets to the total number of pairs in the gold standard (cf. formula 5). Precision and recall are finally combined to produce the harmonic mean (cf. formula 6).

$$P = \frac{|F(K) \cap F(S)|}{|F(K)|} \tag{4}$$

$$R = \frac{|F(K) \cap F(S)|}{|F(S)|} \tag{5}$$

$$FS = \frac{2 \cdot P \cdot R}{P + R} \tag{6}$$

4.2. Baselines

The obtained results are compared to three baselines following Agirre and Soroa (2007):

1. a 'one cluster per word' baseline (1c1word), which groups all paraphrases of the target word into a single cluster (corresponds to most frequent sense (MFS));

2. a 'one cluster per paraphrase' baseline (1c1par), where each paraphrase forms a distinct cluster;

3. a 'random' baseline, where it is randomly defined whether a paraphrase pair is semantically related or not and graph edges are added accordingly. The reported random baseline figures are averages over 5 runs.

[3]These metrics may not be optimal where the clusters overlap. Our component clusters do not contain any overlapping clusters. For the cliques only 8.7% of the paraphrases belong to more than one cluster so overlap is not generally an issue. In future we will investigate other clustering metrics such as proposed by Amigo et al. (2009) and Jurgens and Klaptafis (2013) which deal with overlap.

Clustering	#lemmas with higher F-score			
	Results	**1c1word**	**Neither**	**Total**
Cliques	25	59	49	133
Components	22	36	75	133

Table 2: Number of Lemmas having higher F-score: automatic clustering vs 1c1word.

According to Manandhar et al. (2010), the MFS (1c1word) baseline has a V-Measure equal to 0 since by definition its completeness is 1 and homogeneity is 0. Inversely, the 1c1par baseline would have completeness 0 and homogeneity 1, so the V-Measure would still be 0. As shown in the next section, this is not always the case in our results because our gold data sometimes contains 1 reference cluster, which means that homogeneity for the MFS baseline can be 1 for some target words.

4.3. Results

Table 1 contains the evaluation results. Our clusters, especially the coarse-grained ones (components), outperform the 1c1par baseline on both paired F-Score and V-measure. This shows that the clusters encode correct semantic relations between the paraphrases, which 1c1par puts in separate clusters. Moreover, the clusters perform better than the 1c1word (MFS) baseline on V-measure. The V-measure does tend to be biased towards solutions with more clusters (Manandhar et al., 2010), however that cannot be the only factor since our method produces comparable or superior results to 1c1par which generates the maximum number of clusters per lemma.

According to F-score, 1c1word is stronger which is not surprising (none of the systems participating in the SemEval-2010 Word Sense Induction & Disambiguation task outperformed this baseline in the paired F-score evaluation (Manandhar et al., 2010)). Of course the success of the 1c1word baseline is partly due to the fact that a large number of lemmas (70 out of 133)[4] in the data used in our study have only 1 cluster in the reference as well as the inherent noise in automatic clustering. Nevertheless, there are still 63 lemmas where clustering would in theory be

[4]Note that 133 LexSub lemmas are retained after the cluster projection process.

advantageous, and when we look at comparable F-score performance between the algorithm and 1c1word on a lemma by lemma basis (see Table 2), we note that in fact there is a sizable proportion where automatic clustering outperforms this baseline as well as many where the results are even (neither does better). Finally, our clusters fit the gold data better than the Random clustering does. We note though that Random does relatively well because of the low number of paraphrases and clusters that does not allow for much variation in different clustering solutions.

5. Conclusion

In this study, we have analysed the semantics of paraphrases derived from bilingual parallel corpora by the pivot method, which constitute a valuable resource in multilingual NLP applications. Our results demonstrate that the automatically extracted paraphrases are not always semantically related and highlight the need for an additional semantic analysis stage in applications exploiting this type of resource. A disambiguation module accounting for the sense distinctions that exist in this dataset would help to avoid erroneous matchings and substitutions that might result from using pivot paraphrases in their raw form. Naturally, success will depend on the accuracy of the disambiguation but our study demonstrates that ambiguity exists in paraphrases derived by the pivot method, a finding that certainly merits further exploration.

6. Acknowledgments

This work was partly funded by the French National Research Agency (ANR) project TransRead.

7. References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic, June. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Inf. Retr.*, 12(4):461–486, August.

Marianna Apidianaki and Yifan He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT-10)*, pages 219–226, Paris, France.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 534–542, Singapore, August. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Mohit Bansal, John DeNero, and Dekang Lin. 2012. Unsupervised Translation Sense Clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Montréal, Canada, June. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-sequence Alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 16–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France, July. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June. Association for Computational Linguistics.

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii, October. Association for Computational Linguistics.

Kenneth W. Church and Patrick Hanks. 1991. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16:22–29.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(4).

Helge Dyvik. 1998. Translations as semantic mirrors: from parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton, UK.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003.

Statistical Phrase-Based Translation. In *Proceedings of 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pages 48–54, Edmonton, Canada. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using Paraphrases for Parameter Tuning in Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation*, Prague, Czech Republic, June. Association for Computational Linguistics.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden, July. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June. Association for Computational Linguistics.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Matthew G. Snover, Nitin Madnani, Bonnie J. Dorr, and Richard M. Schwartz. 2010. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia, July. Association for Computational Linguistics.