SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity

Eneko Agirre  
University of the Basque Country  
Donostia, 20018, Basque Country  
e.agirre@ehu.es

Daniel Cer  
Stanford University  
Stanford, CA 94305, USA  
danielcer@stanford.edu

Mona Diab  
Center for Computational Learning Systems  
Columbia University  
mdiab@ccls.columbia.edu

Aitor Gonzalez-Agirre  
University of the Basque Country  
Donostia, 20018, Basque Country  
agonzalez278@ikasle.ehu.es

Abstract

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two texts. This paper presents the results of the STS pilot task in Semeval. The training data contained 2000 sentence pairs from previously existing paraphrase datasets and machine translation evaluation resources. The test data also comprised 2000 sentences pairs for those datasets, plus two surprise datasets with 400 pairs from a different machine translation evaluation corpus and 750 pairs from a lexical resource mapping exercise. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk, with high Pearson correlation scores, around 90%. 35 teams participated in the task, submitting 88 runs. The best results scored a Pearson correlation >80%, well above a simple lexical baseline that only scored a 31% correlation. This pilot task opens an exciting way ahead, although there are still open issues, specially the evaluation metric.

1 Introduction

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two sentences. STS is related to both Textual Entailment (TE) and Paraphrase (PARA). STS is more directly applicable in a number of NLP tasks than TE and PARA such as Machine Translation and evaluation, Summarization, Machine Reading, Deep Question Answering, etc. STS differs from TE in as much as it assumes symmetric graded equivalence between the pair of textual snippets. In the case of TE the equivalence is directional, e.g. a car is a vehicle, but a vehicle is not necessarily a car. Additionally, STS differs from both TE and PARA in that, rather than being a binary yes/no decision (e.g. a vehicle is not a car), STS incorporates the notion of graded semantic similarity (e.g. a vehicle and a car are more similar than a wave and a car).

STS provides a unified framework that allows for an extrinsic evaluation of multiple semantic components that otherwise have tended to be evaluated independently and without broad characterization of their impact on NLP applications. Such components include word sense disambiguation and induction, lexical substitution, semantic role labeling, multi-word expression detection and handling, anaphora and coreference resolution, time and date resolution, named-entity handling, underspecification, hedging, semantic scoping and discourse analysis. Though not in the scope of the current pilot task, we plan to explore building an open source toolkit for integrating and applying diverse linguistic analysis modules to the STS task.

While the characterization of STS is still preliminary, we observed that there was no comparable existing dataset extensively annotated for pairwise sentence similarity. We approached the construction of the first STS dataset with the following goals: (1) To set a definition of STS as a graded notion which can be easily communicated to non-expert annotators beyond the likert-scale; (2) To gather a substantial amount of sentence pairs from diverse datasets, and to annotate them with high quality; (3) To explore evaluation measures for STS; (4) To explore the relation of STS to PARA and Machine Translation Evaluation exercises.
In the next section we present the various sources of the STS data and the annotation procedure used. Section 4 investigates the evaluation of STS systems. Section 5 summarizes the resources and tools used by participant systems. Finally, Section 6 draws the conclusions.

2 Source Datasets

Datasets for STS are scarce. Existing datasets include (Li et al., 2006) and (Lee et al., 2005). The first dataset includes 65 sentence pairs which correspond to the dictionary definitions for the 65 word pairs in Similarity (Rubenstein and Goodenough, 1965). The authors asked human informants to assess the meaning of the sentence pairs on a scale from 0.0 (minimum similarity) to 4.0 (maximum similarity). While the dataset is very relevant to STS, it is too small to train, develop and test typical machine learning based systems. The second dataset comprises 50 documents on news, ranging from 51 to 126 words. Subjects were asked to judge the similarity of document pairs on a five-point scale (with 1.0 indicating “highly unrelated” and 5.0 indicating “highly related”). This second dataset comprises a larger number of document pairs, but it goes beyond sentence similarity into textual similarity.

When constructing our datasets, gathering naturally occurring pairs of sentences with different degrees of semantic equivalence was a challenge in itself. If we took pairs of sentences at random, the vast majority of them would be totally unrelated, and only a very small fragment would show some sort of semantic equivalence. Accordingly, we investigated reusing a collection of existing datasets from tasks related to STS.

We first studied the pairs of text from the Recognizing TE challenge. The first editions of the challenge included pairs of sentences as the following:

T: The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.
H: Jill Carroll was abducted in Iraq.

The first sentence is the text, and the second is the hypothesis. The organizers of the challenge annotated several pairs with a binary tag, indicating whether the hypothesis could be entailed from the text. Although these pairs of text are interesting we decided to discard them from this pilot because the length of the hypothesis was typically much shorter than the text, and we did not want to bias the STS task in this respect. We may, however, explore using TE pairs for STS in the future.

Microsoft Research (MSR) has pioneered the acquisition of paraphrases with two manually annotated datasets. The first, called MSR Paraphrase (MSRpar for short) has been widely used to evaluate text similarity algorithms. It contains 5801 pairs of sentences gleaned over a period of 18 months from thousands of news sources on the web (Dolan et al., 2004). 67% of the pairs were tagged as paraphrases. The inter annotator agreement is between 82% and 84%. Complete meaning equivalence is not required, and the annotation guidelines allowed for some relaxation. The pairs which were annotated as not being paraphrases ranged from completely unrelated semantically, to partially overlapping, to those that were almost-but-not-quite semantically equivalent. In this sense our graded annotations enrich the dataset with more nuanced tags, as we will see in the following section. We followed the original split of 70% for training and 30% for testing. A sample pair from the dataset follows:

The Senate Select Committee on Intelligence is preparing a blistering report on prewar intelligence on Iraq.

American intelligence leading up to the war on Iraq will be criticized by a powerful US Congressional committee due to report soon, officials said today.

In order to construct a dataset which would reflect a uniform distribution of similarity ranges, we sampled the MSRpar dataset at certain ranks of string similarity. We used the implementation readily accessible at CPAN\(^1\) of a well-known metric (Ukkonen, 1985). We sampled equal numbers of pairs from five bands of similarity in the [0.4 .. 0.8] range separately from the paraphrase and non-paraphrase pairs. We sampled 1500 pairs overall, which we split 50% for training and 50% for testing.

The second dataset from MSR is the MSR Video Paraphrase Corpus (MSRvid for short). The authors showed brief video segments to Annotators from Amazon Mechanical Turk (AMT) and were asked

\(^1\)http://search.cpan.org/˜mlehmann/ String-Similarity-1.04/Similarity.pm
Given the strong connection between STS systems and Machine Translation evaluation metrics, we also sampled pairs of segments that had been part of human evaluation exercises. Those pairs included a reference translation and a automatic Machine Translation system submission, as follows:

The only instance in which no tax is levied is when the supplier is in a non-EU country and the recipient is in a Member State of the EU.

The only case for which no tax is still perceived “is an example of supply in the European Community from a third country.

We selected pairs from the translation shared task of the 2007 and 2008 ACL Workshops on Statistical Machine Translation (WMT) (Callison-Burch et al., 2007; Callison-Burch et al., 2008). For consistency, we only used French to English system submissions. The training data includes all of the Europarl human ranked fr-en system submissions from WMT 2007, with each machine translation being paired with the correct reference translation. This resulted in 729 unique training pairs. The test data is comprised of all Europarl human evaluated fr-en pairs from WMT 2008 that contain 16 white space delimited tokens or less.

In addition, we selected two other datasets that were used as out-of-domain testing. One of them comprised of all the human ranked fr-en system submissions from the WMT 2007 news conversation test set, resulting in 351 unique system reference pairs. The second set is radically different as it comprised 750 pairs of glosses from OntoNotes 4.0 (Hovy et al., 2006) and WordNet 3.1 (Fellbaum, 1998) senses. The mapping of the senses of both resources comprised 110K sense pairs. The similarity between the sense pairs was generated using simple word overlap. 50% of the pairs were sampled from senses which were deemed as equivalent senses, the rest from senses which did not map to one another.

3 Annotation

In this first dataset we defined a straightforward likert scale ranging from 5 to 0, but we decided to provide definitions for each value in the scale (cf. Figure 2). We first did pilot annotations of 200 pairs se-

\[\text{Figure 2: Definition and instructions for annotation}\]

\[
\text{Compare Two Similar Sentences}\\
\text{Scene how similar two sentences are to each other according to the following scale.}\\
\text{The sentences are:}\\
(5) Completely equivalent, as they mean the same thing.\\
(4) Mostly equivalent, but some unimportant details differ.\\
(3) Roughly equivalent, but some important information differs/missing.\\
(2) Not equivalent, but share some details.\\
(1) Not equivalent, but are on the same topic.\\
(0) On different topics.\\
Select a similarity rating for each sentence pair below:\n\]

\[\text{Figure 2: Definition and instructions for annotation}\]

\[\text{Given the strong connection between STS systems and Machine Translation evaluation metrics, we also sampled pairs of segments that had been part of human evaluation exercises. Those pairs included a reference translation and an automatic Machine Translation system submission, as follows:}\]

\[\text{The only instance in which no tax is levied is when the supplier is in a non-EU country and the recipient is in a Member State of the EU.}\]

\[\text{The only case for which no tax is still perceived "is an example of supply in the European Community from a third country.}\]

\[\text{We selected pairs from the translation shared task of the 2007 and 2008 ACL Workshops on Statistical Machine Translation (WMT) (Callison-Burch et al., 2007; Callison-Burch et al., 2008). For consistency, we only used French to English system submissions. The training data includes all of the Europarl human ranked fr-en system submissions from WMT 2007, with each machine translation being paired with the correct reference translation. This resulted in 729 unique training pairs. The test data is comprised of all Europarl human evaluated fr-en pairs from WMT 2008 that contain 16 white space delimited tokens or less.}\]

\[\text{In addition, we selected two other datasets that were used as out-of-domain testing. One of them comprised of all the human ranked fr-en system submissions from the WMT 2007 news conversation test set, resulting in 351 unique system reference pairs. The second set is radically different as it comprised 750 pairs of glosses from OntoNotes 4.0 (Hovy et al., 2006) and WordNet 3.1 (Fellbaum, 1998) senses. The mapping of the senses of both resources comprised 110K sense pairs. The similarity between the sense pairs was generated using simple word overlap. 50% of the pairs were sampled from senses which were deemed as equivalent senses, the rest from senses which did not map to one another.}\]

\[\text{3 Annotation}\]

\[\text{In this first dataset we defined a straightforward likert scale ranging from 5 to 0, but we decided to provide definitions for each value in the scale (cf. Figure 2). We first did pilot annotations of 200 pairs se-}\]

\[\text{2At the time of the shared task, this data set contained duplicates resulting in 399 sentence pairs.}\]
lected at random from the three main datasets in the training set. We did the annotation, and the pairwise Pearson ranged from 84% to 87% among ourselves. The agreement of each annotator with the average scores of the other was between 87% and 89%.

In the future, we would like to explore whether the definitions improve the consistency of the tagging with respect to a likert scale without definitions. Note also that in the assessment of the quality and evaluation of the systems performances, we just took the resulting SS scores and their averages. Using the qualitative descriptions for each score in analysis and evaluation is left for future work.

Given the good results of the pilot we decided to deploy the task in Amazon Mechanical Turk (AMT) in order to crowd source the annotation task. The turkers were required to have achieved a 95% of approval rating in their previous HITs, and had to pass a qualification task which included 6 example pairs. Each HIT included 5 pairs of sentences, and was paid at 0.20$ each. We collected 5 annotations per HIT. In the latest data collection, each HIT required 114.9 second for completion.

In order to ensure the quality, we also performed post-hoc validation. Each HIT contained one pair from our pilot. After the tagging was completed we checked the correlation of each individual turker with our scores, and removed annotations of turkers which had low correlations (below 50%). Given the high quality of the annotations among the turkers, we could alternatively use the correlation between the turkers itself to detect poor quality annotators.

### 4 Systems Evaluation

Given two sentences, s1 and s2, an STS system would need to return a similarity score. Participants can also provide a confidence score indicating their confidence level for the result returned for each pair, but this confidence is not used for the main results. The output of the systems performance is evaluated using the Pearson product-moment correlation coefficient between the system scores and the human scores, as customary in text similarity (Rubenstein and Goodenough, 1965). We calculated Pearson for each evaluation dataset separately.

In order to have a single Pearson measure for each system we concatenated the gold standard (and system outputs) for all 5 datasets into a single gold standard file (and single system output). The first version of the results were published using this method, but the overall score did not correspond well to the individual scores in the datasets, and participants proposed two additional evaluation metrics, both of them based on Pearson correlation. The organizers of the task decided that it was more informative, and on the benefit of the community, to also adopt those evaluation metrics, and the idea of having a single main evaluation metric was dropped. This decision was not without controversy, but the organizers gave more priority to openness and inclusiveness and to the involvement of participants. The final result table thus included three evaluation metrics. For the future we plan to analyze the evaluation metrics, including non-parametric metrics like Spearman.

#### 4.1 Evaluation metrics

The first evaluation metric is the Pearson correlation for the concatenation of all five datasets, as described above. We will use overall Pearson or simply ALL to refer to this measure.

The second evaluation metric normalizes the output for each dataset separately, using the linear least squares method. We concatenated the system results for five datasets and then computed a single Pearson correlation. Given \( Y = \{y_i\} \) and \( X = \{x_i\} \) (the gold standard scores and the system scores, respectively), we transform the system scores into \( X' = \{x_i'\} \) in order to minimize the squared error \( \sum_i (y_i - x_i')^2 \). The linear transformation is given by \( x_i' = x_i \beta_1 + \beta_2 \), where \( \beta_1 \) and \( \beta_2 \) are found analytically. We refer to this measure as Normalized Pearson or simply ALLnorm. This metric was suggested by one of the participants, Sergio Jimenez.

The third evaluation metric is the weighted mean of the Pearson correlations on individual datasets. The Pearson returned for each dataset is weighted according to the number of sentence pairs in that dataset. Given \( r_i \) the five Pearson scores for each dataset, and \( n_i \) the number of pairs in each dataset, the weighted mean is given as \( \frac{\sum_{i=1..5} (r_i \times n_i)}{\sum_{i=1..5} n_i} \). We refer to this measure as weighted mean of Pearson or Mean for short.

#### 4.2 Using confidence scores

Participants were allowed to include a confidence score between 1 and 100 for each of their scores. We used weighted Pearson to use those confidence
scores. Table 2 includes the list of systems which provided a non-uniform confidence. The results show that some systems were able to improve their correlation, showing promise for the usefulness of confidence in applications.

### 4.3 The Baseline System

We produced scores using a simple word overlap baseline system. We tokenized the input sentences splitting at white spaces, and then represented each sentence as a vector in the multidimensional token space. Each dimension had 1 if the token was present in the sentence, 0 otherwise. Similarity of vectors was computed using cosine similarity.

We also run a random baseline several times, yielding close to 0 correlations in all datasets, as expected. We will refer to the random baseline again in Section 4.5.

### 4.4 Participation

Participants could send a maximum of three system runs. After downloading the test datasets, they had a maximum of 120 hours to upload the results. 35 teams participated, submitting 88 system runs (cf. first column of Table 1). Due to lack of space we can’t detail the full names of authors and institutions that participated. The interested reader can use the name of the runs to find the relevant paper in these proceedings.

There were several issues in the submissions. The submission software did not ensure that the naming conventions were appropriately used, and this caused some submissions to be missed, and in two cases the results were wrongly assigned. Some participants returned Not-a-Number as a score, and the organizers had to request whether those where to be taken as a 0 or as a 5.

Finally, one team submitted past the 120 hour deadline and some teams sent missing files after the deadline. All those are explicitly marked in Table 1. The teams that included one of the organizers are also explicitly marked. We want to stress that in these teams the organizers did not allow the developers of the system to access any data or information which was not available for the rest of participants. One exception is weiwei, as they generated the 110K OntoNotes-WordNet dataset from which the other organizers sampled the surprise data set.

After the submission deadline expired, the organizers published the gold standard in the task website, in order to ensure a transparent evaluation process.

### 4.5 Results

Table 1 shows the results for each run in alphabetic order. Each result is followed by the rank of the system according to the given evaluation measure. To the right, the Pearson score for each dataset is given. In boldface, the three best results in each column.

First of all we want to stress that the large majority of the systems are well above the simple baseline, although the baseline would rank 70 on the Mean measure, improving over 19 runs.

The correlation for the non-MT datasets were really high: the highest correlation was obtained was for MSRvid (0.88 $r$), followed by MSRpar (0.73 $r$) and On-WN (0.73 $r$). The results for the MT evaluation data are lower, (0.57 $r$) for SMT-eur and (0.61 $r$) for SMT-News. The simple token overlap baseline, on the contrary, obtained the highest results for On-WN (0.59 $r$), with (0.43 $r$) on MSRpar and (0.40 $r$) on MSRvid. The results for MT evaluation data are also reversed, with (0.40 $r$) for SMT-eur and (0.45 $r$) for SMT-News.

The ALLnorm measure yields the highest correlations. This comes at no surprise, as it involves a normalization which transforms the system outputs using the gold standard. In fact, a random baseline which gets Pearson correlations close to 0 in all datasets would attain Pearson of 0.5891.

Although not included in the results table for lack of space, we also performed an analysis of confidence intervals. For instance, the best run according to ALL ($r = .8239$) has a 95% confidence interval of [.8123,.8349] and the second a confidence interval of [.8016,.8254], meaning that the differences are not statistically different.

### 5 Tools and resources used

The organizers asked participants to submit a description file, special emphasis on the tools and resources that they used. Table 3 shows in a simpli-
| Run              | ALL Rank | ALLnrm Rank | Mean Rank | MSRpar | MSRvid | SMT-eur | On-WN | SMT-news |
|------------------|----------|-------------|-----------|--------|--------|---------|-------|---------|
| 00-baseline/task6-baseline | 0.3110   | 87          | 0.6732    | 0.85   | 0.56   | 0.67    | 0.85  | 0.56    |

Note: ∗system submitted past the 120 hour window, †team involving one of the organizers.}

Table 1: The first row corresponds to the baseline. All for overall Pearson. All mean for Pearson after normalization and Mean for mean of Pearsons. We also show the ranks for each measure. Rightmost columns show Pearson for each individual dataset. Note: ∗system submitted past the 120 hour window, †team involving one of the organizers.
6 Conclusions and Future Work

This paper presents the SemEval 2012 pilot evaluation exercise on Semantic Textual Similarity. A simple definition of STS beyond the likert-scale was set up, and a wealth of annotated data was produced. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk. The dataset includes 1500 sentence pairs from MSRpar and MSRvid (each), ca. 1500 pairs from WMT, and 750 sentence pairs from a mapping between OntoNotes and WordNet senses. The correlation between non-expert annotators and annotations from the authors is very high, showing the high quality of the dataset. The dataset was split 50% as train and test, with the exception of the surprise test datasets: a subset of WMT from a different domain and the OntoNotes-WordNet mapping. All datasets are publicly available.\(^5\)

The exercise was very successful in participation and results. 35 teams participated, submitting 88 runs. The best results scored a Pearson correlation over 80%, well beyond a simple lexical baseline with 31% of correlation. The metric for evaluation was not completely satisfactory, and three evaluation metrics were finally published. We discuss the shortcomings of those measures.

There are several tasks ahead in order to make STS a mature field. The first is to find a satisfactory evaluation metric. The second is to analyze the definition of the task itself, with a thorough analysis of the definitions in the likert scale.

We would also like to analyze the relation between the STS scores and the paraphrase judgements in MSR, as well as the human evaluations in WMT. Finally, we would also like to set up an open framework where NLP components and similarity algorithms can be combined by the community. All in all, we would like this dataset to be the focus of the community working on algorithmic approaches for semantic processing and inference at large.

Acknowledgements

We would like to thank all participants, specially (in alphabetic order) Yoan Gutierrez, Michael Heilman, Sergio Jimenez, Nitin Madnani, Diana McCarthy and Sruirian Satpathy for their contributions on evaluation metrics. Eneko Agirre was partially funded by the

---

\(^5\)http://www.cs.york.ac.uk/semeval-2012/task6/

---

Table 2: Results according to weighted correlation for the systems that provided non-uniform confidence alongside their scores.

| Run                          | ALL | ALLW | MSRpar | MSRRsp | MSRRvid | SMT-eur | SMT-news | On-WN | On-WNw | SMT-news | SMT-newsW |
|------------------------------|-----|-------|--------|--------|---------|---------|----------|-------|--------|----------|-----------|
| davidjuscdal/task6-IRIT-pg1  | .428 | .4946 | .4295  | .4682  | .6125   | .6593   | .4952    | .5271  | .5387  | .5574    | .5614      |
| davidjuscdal/task6-IRIT-pg3  | .4813| .5503 | .4171  | .4033  | .6728   | .7048   | .5179    | .5329  | .5526  | .5950    | .3693      |
| davidjuscdal/task6-IRIT-wu    | .4064| .4682 | .4326  | .4035  | .5633   | .6253   | .4585    | .5138  | .5317  | .5189    | .5480      |
| enrique/task6-UNED-H34measures| .4381| .2615 | .5328  | .4494  | .5788   | .4913   | .4785    | .4660  | .6692  | .6440    | .4465      |
| enrique/task6-UNED-HallMeasures| .2791| .2082 | .3861  | .3802  | .2570   | .2343   | .4086    | .4321  | .6006  | .5947    | .5305      |
| enrique/task6-UNED-SPINIST   | .6800| .3754 | .5166  | .5082  | .6083   | .5598   | .4681    | .6442  | .5761  | .4753    | .4143      |
| parthapakray/task6-RU_NLP_Semantic_Syntactic_Approach | .3880| .3636 | .3427  | .3498  | .3549   | .3353   | .4271    | .3989  | .5298  | .4619    | .4034      |
| tiantianzhu7/task6-tiantianzhu7-1 | .5533| .5442 | .4164  | .4341  | .5630   | .5630   | .2083    | .4220  | .5031  | .3745    | .3536      |
| tiantianzhu7/task6-tiantianzhu7-2 | .4157| .5249 | .4260  | .4340  | .3628   | .5758   | .1546    | .4776  | .4552  | .4926    | .3923      |
| tiantianzhu7/task6-tiantianzhu7-3 | .4446| .5229 | .3411  | .3611  | .5946   | .5899   | .1868    | .4769  | .4029  | .4365    | .3823      |
| Acronym | Description | Distributional thesaurus | Monolingual corpora | Multilingual corpora | Tables of paraphrases | Wikipedia | Other |
|---------|-------------|--------------------------|---------------------|---------------------|----------------------|----------|-------|
| DLSI-MultiSem | | | | | | | |
| SMT | | | | | | | |
| TA-STAT | | | | | | | |
| Alignment | | | | | | | |
| Other | | | | | | | |

**Table 3: Resources and tools used by the systems that submitted a description file. Leftmost columns correspond to the resources, and rightmost to tools, in alphabetic order.**
European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082 (PATHS project) and the Ministry of Economy under grant TIN2009-14715-C04-01 (KNOW2 project). Daniel Cer gratefully acknowledges the support of the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181 and the support of the DARPA Broad Operational Language Translation (BOLT) program through IBM. The STS annotations were funded by an extension to DARPA GALE subcontract to IBM # W0853748 4911021461.0 to Mona Diab. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, StatMT ’07, pages 136–158.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In Proceedings of the Third Workshop on Statistical Machine Translation, StatMT ’08, pages 70–106.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meetings of the Association for Computational Linguistics (ACL).

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In COLING 04: Proceedings of the 20th international conference on Computational Linguistics, page 350.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL.

Michael D. Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In Proceedings of the 27th Annual Conference of the Cognitive Science Society, pages 1254–1259, Mahwah, NJ.

Y. Li, D. McLean, Z. A. Bandar, J. D. O’Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8):1138–1150, August.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. Commun. ACM, 8(10):627–633, October.

E. Ukkonen. 1985. Algorithms for approximate string matching. Information and Control, 64:110–118.

393