Crowdsourcing Word Sense Definition

Anna Rumshisky[†*]
[†] Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA
arum@csail.mit.edu
[*]Department of Computer Science
Brandeis University
Waltham, MA

Abstract

In this paper, we propose a crowdsourcing methodology for a single-step construction of both an empirically-derived sense inventory and the corresponding sense-annotated corpus. The methodology taps the intuitions of non-expert native speakers to create an expert-quality resource, and natively lends itself to supplementing such a resource with additional information about the structure and reliability of the produced sense inventories. The resulting resource will provide several ways to empirically measure distances between related word senses, and will explicitly address the question of fuzzy boundaries between them.

1 Introduction

A number of recent initiatives has focused on creating sense-annotated gold standards for word sense disambiguation and induction algorithms. However, such work has frequently come under criticism over the lack of a satisfactory set of standards for creating consistent, task-independent sense inventories. More systematic efforts to replace ad hoc lexicographic procedures for sense inventory construction have often focused on working with existing sense inventories, attempting to resolve the specific associated problems (e.g. sense granularity, overlapping senses, etc.) Methodologically, defining a robust procedure for sense definition has remained an elusive task.

In this paper, we propose a method for creating a sense inventory from scratch for any polysemous word, simultaneously with the corresponding sense-annotated lexical sample. The methodology we propose explicitly addresses the question of related word senses and fuzzy boundaries between them, without trying to establish hard divisions where empirically there are none.

The proposed method uses Amazon's Mechanical Turk for sense annotation. Over the last several of years, Mechanical Turk, introduced by Amazon as "artificial artificial intelligence", has been used successfully for a number of NLP tasks, including robust evaluation of machine translation systems by reading comprehension (Callison-Burch, 2009), and other tasks explored in the recent NAACL workshop (Callison-Burch and Dredze, 2010b). Mechanical Turk has also been used to create labeled data sets for word sense disambiguation (Snow et al., 2008) and even to modify sense inventories. But the original sense inventory construction has always been left to the experts. In contrast, in the annotation method we describe, the expert is eliminated from the annotation process. As has been the case with using Mechanical Turk for other NLP tasks, the proposed annotation is quite inexpensive and can be done very quickly, while maintaining expert-level annotation quality.

The resulting resource will produce several ways to empirically measure distances between senses, and should help to address some open research questions regarding word sense perceptions by native speakers. We describe a set of pilot annotation studies needed to ensure reliability of this methodology and test the proposed quality control mechanisms.

The outcome will be a lexicon where sense inventories are represented as clusters of instances, and an explicit quantitative representation of sense con-

sistency, distance between senses, and sense overlap is associated with the senses for each word. The goal is to provide a more accurate representation the way speakers of a language conceptualize senses, which can be used for training and testing of the automated WSD systems, as well as to automatically induce semantic and syntactic context patterns that represent usage norms and permit native speakers to perform sense disambiguation.

2 The Problem of Sense Definition

The quality of the annotated corpora depends directly on the selected sense inventory, so, for example, SemCor (Landes et al., 1998), which used WordNet synsets, inherited all the associated problems, including using senses that are too fine-grained and in many cases poorly distinguished. At the Senseval competitions (Mihalcea et al., 2004; Snyder and Palmer, 2004; Preiss and Yarowsky, 2001), the choice of a sense inventory also frequently presented problems, spurring the efforts to create coarser-grained sense inventories (Navigli, 2006; Hovy et al., 2006; Palmer et al., 2007). Inventories derived from WordNet by using small-scale corpus analysis and by automatic mapping to top entries in Oxford Dictionary of English were used in the recent workshops on semantic evaluation, including Semeval-2007 and Semeval-2010 (Agirre et al., 2007; Erk and Strapparava, 2010).

Several current resource-oriented projects attempt to formalize the procedure of creating a sense inventory. FrameNet (Ruppenhofer et al., 2006) attempts to organize lexical information in terms of script-like semantic frames, with semantic and syntactic combinatorial possibilities specified for each frame-evoking lexical unit (word/sense pairing). Corpus Pattern Analysis (CPA) (Hanks and Pustejovsky, 2005) attempts to catalog norms of usage for individual words, specifying them in terms of context patterns. Other large-scale resource-building projects also use corpus analysis techniques. In PropBank (Palmer et al., 2005), verb senses were defined based on their use in Wall Street Journal corpus and specified in terms of framesets which consist of a set of semantic roles for the arguments of a particular sense. In the OntoNotes project (Hovy et al., 2006), annotators use small-scale corpus analysis to create sense inventories derived by grouping together WordNet senses, with the procedure restricted to maintain 90% inter-annotator agreement.

Importantly, most standard WSD resources contain no information about the clarity of distinctions between different senses in the sense inventory. For example, OntoNotes, which was used for evaluation in the word sense disambiguation and sense induction tasks in the latest SemEval competitions contains no information about sense hierarchy, related senses, or difficulty and consistency of a given set of senses.

3 Characteristics of the Proposed Lexical Resource

The lexical resource we propose to build is a sense-disambiguated lexicon which will consist of an empirically-derived sense inventory for each word in the language, and a sense-tagged corpus annotated with the derived inventories. The resource will be assembled from "the ground up" using the intuitions of non-expert native speakers about the similarity between different uses of the same word. Each sense will be represented as a cluster of instances grouped together in annotation. The following information will be associated with each sense cluster:

1. Consistency rating for each sense cluster, including several of the following measures:
   - *Annotator agreement*, using the inter-annotator agreement measures for the sense cluster (e.g. Fleiss' Kappa);
   - *Cluster tightness*, determined from the distributional contextual features associated with instance comprising the cluster;
2. Distances to other sense clusters derived for the same word, using several distance measures, including:
   - *Cluster overlap*, determined from the percentage of instances associated with both clusters;
   - *Translation similarity*, determined as the number existing different lexicalizations in an aligned multilingual parallel corpus, using a measurement methodology similar to Resnik and Yarowsky (1999).

The resource would also include a *Membership rating* for each instance within a given sense cluster, which would represent how typical this example is for the associated sense cluster. The instances whose membership in the cluster was established with minimal disagreement between the annotators, and which do not have multiple sense cluster membership will be designated as the core of the sense cluster. The membership ratings would be based on (1) inter-annotator agreement for that instance (2) distance from the core elements of the cluster.

Presently, the evaluation of automated WSD and WSI systems does not take into account the relative difficulty of sense distinctions made within a given sense inventories. In the proposed resource, for every lexical item, annotator agreement values will be associated with each sense separately, as well as with the full sense inventory for that word, providing an innate measure of disambiguation difficulty for every lexical item.

Given that the fluidity of senses is such a pervasive problem for lexical resources and that it creates severe problems for the usability of the systems trained using these resources, establishing the reliability and consistency of each sense cluster and the "prototypicality" of each example associated with that sense is crucial for any lexical resource. Similarly crucial is the information about the overlap between senses in a sense inventory as well as the similarity between senses. And yet, none of the existing resources contain this information.[1] As a result, the systems trained on sense-tagged corpora using the existing sense inventories attempt to make sense distinctions where empirically no hard division between senses exist. And since the information about consistency and instance typicality is not available, the standard evaluation paradigm currently used in the field for the automated WSD/WSI systems does not take it into account. In contrast, the methodology we propose here lends itself naturally to quantitative analysis needed to explicitly address the question of related word senses and fuzzy boundaries between them.

[1]One notable exception is the sense-based inter-annotator agreement available in OntoNotes.

4 Annotation Methodology

In traditional annotation settings, the quality of annotation directly depends on how well the annotation task is defined. The effects of felicitous or poor task design are greatly amplified when one is targeting untrained non-expert annotators.

Typically for the tasks performed using Mechanical Turk, complex annotation is split into simpler steps. Each step is farmed out to the non-expert annotators employed via Mechanical Turk (henceforth, MTurkers) in a form of a HIT (Human Intelligence Task), a term used to refer to the tasks that are hard to perform automatically, yet very easy to do for humans.

4.1 Prototype-Based Clustering

We propose a simple HIT design intended to imitate the work done by a lexicographer in corpus-based dictionary construction, of the kind used in Corpus Pattern Analysis (CPA, 2009). The task is designed as a sequence of annotation rounds, with each round creating a cluster corresponding to one sense. MTurkers are first given a set of sentences containing the target word, and one sentence that is randomly selected from this set as a target sentence. They are then asked to identify, for each sentence, whether the target word is used in the same way as in the target sentence. If the sense is unclear or it is impossible to tell, they are instructed to pick the "unclear" option. After the first round of annotation is completed, the sentences that are judged as similar to the target sentence by the majority vote are set apart into a separate cluster corresponding to one sense, and excluded from the set used in further rounds. The procedure is repeated with the remaining set, i.e. a new target sentence is selected, and the remaining examples are presented to the annotators. This cycle is repeated until all the remaining examples are classified as "unclear" by the majority vote, or no examples remain.

4.2 Proof-of-Concept Study

A preliminary proof-of-concept study for this task design has been reported on previously (Rumshisky et al., 2009). In that study, the proposed task design was tested on a chosen polysemous verb of medium difficulty. The results were then evaluated against

the groupings created by a professional lexicographer, giving the set-matching F-score of 93.0 and the entropy of the two clustering solutions of 0.3. The example sentences were taken from the CPA verb lexicon for *crush*. Figure 1 shows the first screen displayed to MTurkers for the HIT, with ten examples presented on each screen. Each example was annotated by 5 MTurkers.

The prototype sentences associated with each cluster obtained for the verb *crush* are shown below:

C1 By appointing Majid as Interior Minister, President Saddam placed him in charge of **crushing** the southern rebellion.
C2 The lighter woods such as balsa can be **crushed** with the finger.
C3 This time the defeat of his hopes didn't **crush** him for more than a few days.

Each round took approximately 30 minutes to an hour to complete, depending on the number of sentences in that round. Each set of 10 sentences took on the average 1 minute, and the annotator received $0.03 USD as compensation. The experiment was conducted using 5-way annotation, and the total sum spent was less than $10 USD. It should be noted that in a large-scale annotation effort, the cost of the annotation for a single word will certainly vary depending on the number of senses it has. However, time is less of an issue, since the annotators can work in parallel on many words at the same time.

4.3 Removing Prototype Impact

Prototype-based clustering produces hard clusters, without explicit information about the origin of boundary cases or the potentially overlapping senses. One of the possible alternatives to having instances judged against a single prototype, with multiple iterations, is to have pairs of concordance lines evaluated against each other. This is in effect more realistic, since (1) each sentence is effectively a prototype, and (2) there is no limitation on the types of similarity judgments allowed; "cross-cluster" connections can be retained.

Whether obtained in a prototype-based setup, or in pairs, the obtained data lends itself well to a graph representation. The pairwise judgments induce an undirected graph, in which judgments can be thought of as edges connecting the instance nodes, and interconnected clusters of nodes correspond to the derived sense inventory (cf. Figure 2).

In the pairwise setup, results do not depend on the selection of a prototype sentence, so it provides a natural protection against a single unclear sentence having undue impact on cluster results, and does so without having to introduce an additional step into the annotation process. It also protects against directional similarity evaluation bias. However, one of the disadvantages is the number of judgments required to collect. The prototype-based clustering of N instances requires between $N(N-1)/2$ and $N-1$ judgments (depending on the way instances split between senses), which gives $O(N^2)$ for 1 cluster 1 instance case vs. $O(N)$ for 1 cluster 1 word case. A typical sense inventory has < 10 senses, so that gives us an estimate of about $10N$ judgments to cluster N concordance lines, to be multiplied by the number of annotators for each pair. In order to bypass prototyping, we must allow same/different judgments for every pair of examples. For N examples, this gives $O(N^2)$ judgments, which makes collecting all pair judgments, from multiple annotators, too expensive.

One of the alternatives for reducing the number of judgments is to use a partial graph approximation. The idea behind it is that rather than collecting repeat judgments (multiple annotations) of the same instance, one would collect a random subset of edges from the full graph, and then perform clustering on the obtained sparse graph. Full pairwise annotation will need to be performed on a small cross-section of English vocabulary in order to get an idea of how sparse the judgment graph can be to obtain results comparable to those we obtained with prototype-based clustering using good prototypes.

Some preliminary experiments using Markov Clustering (MCL) on a sparse judgment graph suggest that the number of judgments collected in the proof-of-concept experiment above by Rumshisky et al. (2009) in order to cluster 350 concordance lines would only be sufficient to reliably cluster about 140 concordance lines.

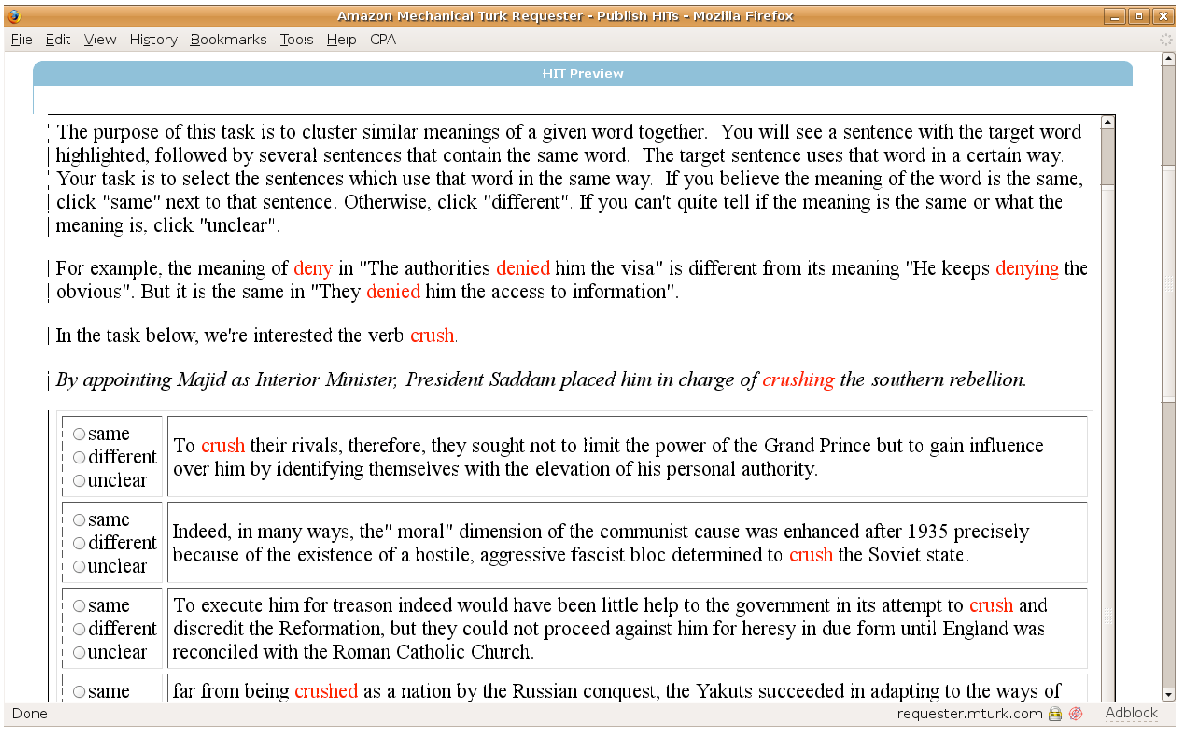

Figure 1: Task interface and instructions for the HIT presented to the non-expert annotators in proof-of-concept experiment.

5 Pilot Annotations

In this section, we outline the pilot studies that need to be conducted prior to applying the described methodology in a large-scale annotation effort. The goal of the pilot studies we propose is to establish the best MTurk annotation practice that would ensure the reliability of obtained results while minimizing the required time and cost of the annotation. The anticipated outcome of these studies is a robust methodology which can be applied to unseen data during the construction of the proposed lexical resource.

5.1 Testing the validity of obtained results

The goal of the first set of studies is to establish the validity of sense groupings obtained using non-expert annotators. We propose to use the procedure outlined in Sec 4 on the data from existing sense-tagged corpora, in particular, OntoNotes, PropBank, NomBank, and CPA.

This group of pilot studies would involve performing prototype-based annotation for a selected set of words representing a cross-section of English vocabulary. A concordance for each selected word will be extracted from the gold standard provided by an expert-tagged sense-annotated corpus. The initial set of selected content words would be evenly split between verbs and nouns. Each group will consist of a set of words with different degrees of polysemy. The lexical items would need to be prioritized according to corpus frequencies, with more frequent words from each group being given preference.

For example, for verbs, a preliminary study done within the framework of the CPA project suggested that out of roughly 6,000 verbs in a language, 30% have one sense, with the rest evenly split between verbs having 2-3 senses and verbs having more than 4 senses. About 20 light verbs have roughly 100 senses each. The chosen lexical sample will therefore need to include low-polysemy verbs, mid-range verbs with 3-10 senses, lighter verbs with 10-20 senses, and several light verbs. Degree of polysemy would need to be obtained from the existing lexical resource used as a gold standard. The annotation procedure could also be tested additionally on a small number of adjectives and adverbs.

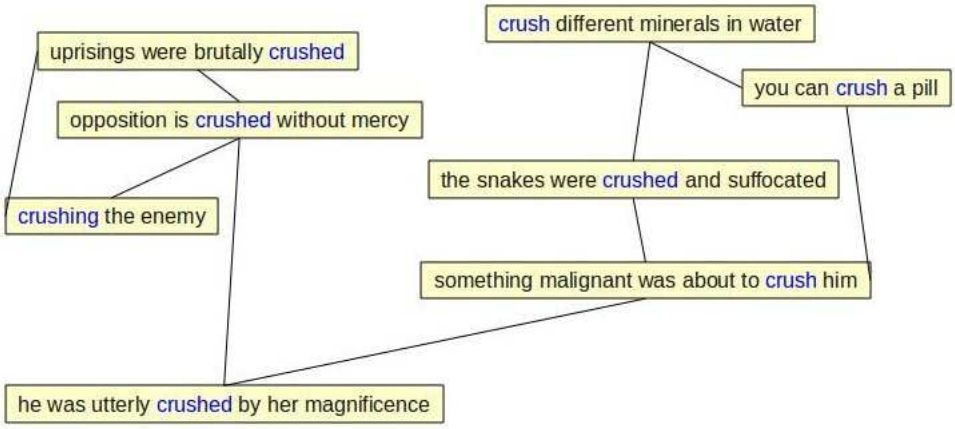

Figure 2: Similarity judgment graph

A smaller subset of the re-annotated data would then need to be annotated using full pairwise annotation. The results of this annotation would need to be used to investigate the quality of the clusters obtained using a partial judgment graph, induced by a subset of collected judgments. The results of both types of annotation could then be used to evaluate different measures of sense consistency and as well as for evaluation of distance between different senses of a lexical item.

5.2 Testing quality control mechanisms

The goal of this set of studies is to establish reliable quality control mechanisms for the annotation. A number of mechanisms for quality control have been proposed for use with Mechanical Turk annotation (Callison-Burch and Dredze, 2010a). We propose to investigate the following mechanisms:

- Multiple annotation. A subset of the data from existing resources would need to be annotated by a larger number of annotators, (e.g. 10 MTurkers. The obtained clustering results would need to be compared to the gold standard data from the existing resource, while varying the number of annotators producing the clustering through majority voting. Results from different subsets of annotators for each subset size would need to be aggregated to evaluate the consistency of annotation for each value. For example, for 3-way annotation, the clusterings obtained from by the majority vote within all possible triads of annotators would be evaluated and the results averaged.

- Checking annotator work against gold standard. Using the same annotated data set, we could investigate the effects of eliminating the annotators performing poorly on the judgments of similarity for the first 50 examples from the gold standard. The judgments of the remaining annotators would need to be aggregated to produce results through a majority vote.

- Checking annotator work against the majority vote. Using a similar approach, we can investigate the effects of eliminating the annotators performing poorly against the majority vote. The data set obtained above would allow us to experiment with different thresholds for eliminating annotators, in each case evaluating the resulting improvement in cluster quality.

- Using prototype-quality control step. We would need to re-annotate a subset of words using an additional step, during which poor quality prototype sentences will be eliminated. This step would be integrated with the main annotation as follows. For each candidate prototype sentence, we would collect the first few similarity judgments from the selected number of annotators. If a certain percentage of judgments are logged as unclear, the sentence is elimi-

nated from the set, and another prototype sentence is selected. We would evaluate the results of this modification, using different thresholds for the number of judgments collected and the percentage of "unclear" ratings.

5.3 Using translation equivalents to compute distances between senses

The goal of this set of studies is to investigate the viability of computing distances between the sense clusters obtained for a given word by using its translation equivalents in other languages. If this methodology proves viable, then the proposed lexical resource can be designed to include some data from multilingual parallel corpora. This would provide both a methodology for measuring relatedness of derived senses and a ready set of translation equivalents for every sense.

Resnik and Yarowsky (1999) used human annotators to produce cross-lingual data in order to measure distances between different senses in a monolingual sense inventory and derive a hierarchy of senses, at different levels of sense granularity. Two methods were tested, where the first one involved asking human translators for the "best" translation for a given polysemous word in a monolingual sense-annotated lexical sample data set. The second method involved asking the human translators, for each pair of examples in the lexical sample, to provide different lexicalizations for the target word, if they existed in their language. The distances between different senses were then determined from the number of languages in which different lexicalizations were preferred (or existed) for different senses of the target word.

In the present project, we propose to obtain similar information by using the English part of a word-aligned multilingual parallel corpus for sense annotation. The degree of cross-lingual lexicalization of the target word in instances associated with different sense classes could then be used to evaluate the distance between these senses. We propose the following to be done as a part of this pilot study. For a selected sample of polysemous words:

- Extract several hundred instances for each word from the English part of a multilingual corpus, such as the Europarl (Koehn, 2005); [2]
- Use the best MTurk annotation procedure as established in Sec 5.2 to cluster the extracted instances;
- Obtain translation equivalents for each instance of the target word using word-alignment produced with Giza++ (Och and Ney, 2000);
- Compute the distances between the obtained clusters by estimating the probability of different lexicalization of the two senses from the word-aligned parallel corpus.

The distances would then be computed using a multilingual cost function similar to the one used by Resnik and Yarowsky (1999), shown in Figure 5.3.

The Europarl corpus contains Indo-European languages (except for Finnish), predominantly of the Romanic and Germanic family. These languages often have parallel sense distinctions. If that proves to be the case, a small additional parallel corpus with the data from other non-European languages would need to be used to supplement the data from Europarl.

6 Conclusion

In this paper, we have presented a proposal for a new annotation strategy for obtaining sense-annotated data WSD/WSI applications, together with the corresponding sense inventories, using non-expert annotators. We have described a set of pilot studies that would need to be conducted prior to applying this strategy in a large-scale annotation effort. We outlined the provisional design of the lexical resource that can be constructed using this strategy, including the native measures for sense consistency and difficulty, distance between related senses, sense overlap, and other parameters necessary for the hierarchical organization of sense inventories.

Acknowledgments

I would like to thank James Pustejovsky and David Tresner-Kirsch for their contributions to this project.

[2] If necessary, the instance set for selected words may be supplemented with the data from other corpora, such as the JRC-Acquis corpus (Steinberger et al., 2006).

$$\text{Cost}(sense_i, sense_j) = \frac{1}{|Languages|} \sum_{L \in Languages} P_L(\text{diff-lexicalization}|sense_i, sense_j)$$

Figure 3: Multilingual cost function for distances between senses.

References

E. Agirre, L. Màrquez, and R. Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. ACL, Prague, Czech Republic, June.

Chris Callison-Burch and Mark Dredze. 2010a. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June. ACL.

Chris Callison-Burch and Mark Dredze, editors. 2010b. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ACL, Los Angeles, June.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*.

CPA. 2009. Corpus Pattern Analysis.

Katrin Erk and Carlo Strapparava, editors. 2010. *Proceedings of the 5th International Workshop on Semantic Evaluation*. ACL, Uppsala, Sweden, July.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. ACL.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5. Citeseer.

S. Landes, C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *Wordnet: an electronic lexical database*. MIT Press, Cambridge (Mass.).

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July. ACL.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia, July. ACL.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Palmer, H. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

J Preiss and D. Yarowsky, editors. 2001. *Proceedings of the Second Int. Workshop on Evaluating WSD Systems (Senseval 2)*. ACL2002/EACL2001.

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–134.

A. Rumshisky, J. Moszkowicz, and M. Verhagen. 2009. The holy grail of sense definition: Creating a sense-disambiguated corpus from scratch. In *Proceedings of 5th International Conference on Generative Approaches to the Lexicon (GL2009)*, Pisa, Italy.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

R. Snow, B. OConnor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fastbut is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*.

B. Snyder and M. Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. ACL.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Arxiv preprint cs/0609058*.