Turk Bootstrap Word Sense Inventory 2.0: A Large-Scale Resource for Lexical Substitution

Chris Biemann
UKP lab, Technische Universität Darmstadt
Hochschulstr. 10, 64289 Darmstadt, Germany
E-mail: biem@cs.tu-darmstadt.de

Abstract

This paper presents the Turk Bootstrap Word Sense Inventory (TWSI) 2.0. This lexical resource, created by a crowdsourcing process using Amazon Mechanical Turk (http://www.mturk.com), encompasses a sense inventory for lexical substitution for 1,012 highly frequent English common nouns. Along with each sense, a large number of sense-annotated occurrences in context are given, as well as a weighted list of substitutions. Sense distinctions are not motivated by lexicographic considerations, but driven by substitutability: two usages belong to the same sense if their substitutions overlap considerably.
After laying out the need for such a resource, the data is characterized in terms of organization and quantity. Then, we briefly describe how this data was used to create a system for lexical substitutions. Training a supervised lexical substitution system on a smaller version of the resource resulted in well over 90% acceptability for lexical substitutions provided by the system. Thus, this resource can be used to set up reliable, enabling technologies for semantic natural language processing (NLP), some of which we discuss briefly.

Keywords: lexical substitution, crowdsourcing, word sense inventory

1. Introduction

Lexical substitution (McCarthy and Navigli, 2007; Sinha et al., 2009) is an enabling task for a number of NLP systems. Being able to reliably substitute words in context for other words facilitates applications like semantic indexing, question answering, document similarity and machine translation.
Attempts to use lexicographic resources like WordNet (Miller et al., 1990) as a source for synonyms – which are possible substitutions – are hampered by the fine-grained sense structure of these resources: While synonyms for monosemous words can be safely substituted in most contexts, ambiguous words have to be sense-disambiguated before being able to supply the correct (of many possible) substitutions. The fine-grained sense structure of WordNet is widely seen as the blocking issue for word sense disambiguation systems, which do not exceed 80% accuracy for WordNet sense distinctions (see Agirre and Edmonds, 2006). One way of mitigating this is to group WordNet senses; another is to rebuild a sense inventory from scratch that does not suffer from these problems – an approach, which was taken for the resource described here.
Word Sense Disambiguation (WSD) is either performed in a knowledge-based way or cast as a supervising machine learning task. Supervised WSD generally reaches higher accuracy, but comes at the cost of annotating a large number of words in context with their sense.
This paper describes a sense-annotated resource: for over 1,000 very frequent nouns in Wikipedia, a sense inventory was created that draws sense distinctions according to common substitutions. Cost was kept relatively low using crowdsourcing, while the number of examples per sense is high, specifically targeting supervised approaches that predict substitutions for an unseen occurrence of a target word according to the similarity of its context to training data.

2. Creation of the TWSI

This section shortly describes the creation of the TWSI. For a more elaborate description of the process, the reader is referred to Biemann and Nygaard (2010), where a previously released subset of this resource is described.
The TWSI was acquired using a bootstrapping cycle that involves three crowdsourcing tasks. In the first task, workers were asked to supply substitutions for a word in its sentence-wide context. Occurrences of the same word are clustered by substitution overlap. This automatic clustering is manually verified by a second crowdsourcing task. One representative occurrence (with its sentence context) per cluster is chosen as a gloss for the sense. In the third task, workers match the meaning of new sentences for the target word with the current set of glosses for this word. If a large number of occurrences can be matched without finding missing senses, the collection process for this word terminates. Otherwise, substitutions for these yet-unmatched occurrences are gathered just like in the first task, and the cycle starts again.
This results in a sense inventory, in which a sense is characterized by a gloss and a list of substitutions, weighted by their frequency. Further, a large number of occurrences are labelled by sense. As a corpus source, a Wikipedia dump from January 2008 was used.

3. Data Format

This section describes the format and exemplifies the data available per target. It can serve as a manual for accessing the TWSI.
The main directory structure contains the following:

- `targets.txt`: File that contains all 1012 target nouns, one per line
- `readme.txt`: text file that explains the structure of the resource
- `license.txt`: The license text (Attribution-ShareAlike 3.0 Unported)
- `doc/`: Subfolder with documentation

- `inventory/`: Subfolder containing the glosses
- `substitutions/`: Subfolder containing the substitutions per sense
- `contexts/`: Subfolder containing sense-labelled occurrences in sentence contexts
- `corpus/`: Subfolder containing sentences and sentence numbers used throughout for reference, as well as sentence source information

Each subfolder contains separate files, that contain the respective data for a single target.
To exemplify the content, let us take a look at some data available for the target "magazine".
In the file `inventory/magazine.proto`, two senses are found, exemplified by two glosses:

1. magazine@@1: *Their first album was released by Columbia Records in 1972 , and they were voted " Best New Band " by Creem magazine* . (magazine++51110955)
2. magazine@@2: *Instead , the film is pulled through the camera solely through the power of camera sprockets until the end , at which point springs or belts in the camera magazine pull the film back to the take - up side* . (magazine++10845213)

Senses are marked by @@<number>. Also, the sentence numbers (here: 51110955 and 10845213) are given, from which we can determine that the first sentence comes from the Wikipedia-article "Eric Bloom" and the second sentence originates from "Camera magazine" – this information is contained in `corpus/wiki_title_sent.txt`.
In `substitutions/magazine.substitutions`, a frequency-pruned list of substitutions is given, yielding the following data (multiplicity in brackets):

1. magazine@@1: publication [42], periodical [32], journal [30], manual [9], gazette [5], newsletter [4], annual [3], digest [3], circular [2]
2. magazine@@2: cartridge [6], clip [5], chamber [3], holder [3], mag [3], ammunition chamber [2], cache [2], loading chamber [2]

The unpruned version is available in the file `substitutions/raw_data/all_substitutions/magazine.substitutions`. The multiplicities reflect the number of times a substitution was provided by the annotators for a target in context that has been assigned to the respective sense. More frequent senses typically have more substitution of higher multiplicities, due to the creation methodology.
When one is not interested in substitutions grouped by sense, but in substitutions for single sentences, `substitutions/raw_data/substitutions_per_sentence/magazine.turkresults` tells us for example that for sentence 10845213 (see above), the following substitutions were given: cartridge holder [1], clip [1], holder[1], film chamber[1], chamber[1], cache [1], cartridge [1], depository [1], loading chamber[1].

Judgments about the difficulty of providing a substitution for the given sentence are also available; these were used to exclude underspecified contexts from the sense inventory construction.
In `contexts/magazine.contexts`, 189 sentences containing magazine in sense magazine@@1 and 5 sentences labelled with magazine@@2 are given, along with a confidence value (number of workers with this judgment divided by total number of judgments). We learn for example that the sentence "*As with any other hi - cap you fill the magazine , then wind the wheel until there ' s a louder click than normal which indicates that the clockwork mechanism is wound tight* ." uses magazine in the second sense, which was agreed upon by all annotators. The occurrences in sentences are marked in -tags.
All data files are organized in columns separated by ta delimiters. Sentence IDs are used throughout, and all refer to the sentences listed in the `corpus` subfolder.

4. Quantitative Characteristics

This section characterizes the TWSI 2.0 quantitatively. A total of 1,012 target nouns are grouped into 2,443 senses, on average 2.41 per target. Figure 1 shows the distribution of the number of senses per target. Due to sense granularity defined by substitution equivalence, the average number of senses is at about a third compared to WordNet, and similar to other coarse-grained inventories, such as OntoNotes (Hovy et al. 2006).

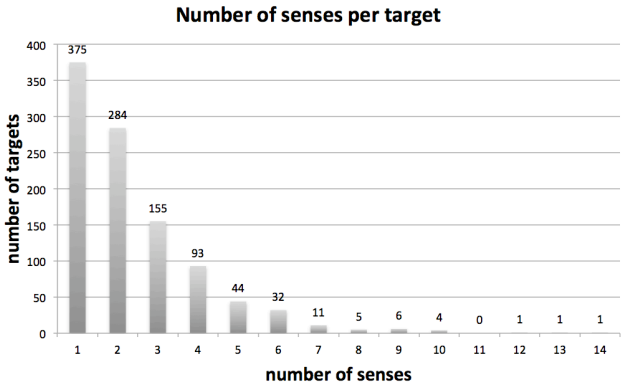

Figure 1: Number of senses per target

The distribution in Figure 1 demonstrates that even for very frequent nouns, a sizeable chunk is monosemous with respect to the corpus and the substitutability criterion. Further, it becomes clear that the creation methodology is able to handle a large number of sense distinctions. For example, the target "stock" is distinguished into the following senses (as given by the most salient substitution): supply, stock theater, reputation, ancestry, barrel stock, share, standard, livestock, raw material, deck/pile, broth.
Figure 2 gives an impression of the richness of the set of substitutions by showing the number of (distinct) substitutions per sense. Most senses received 10 substitutions.

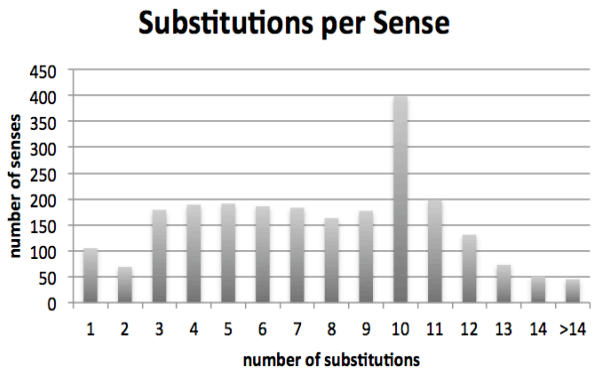

Figure 2: Substitutions per sense

Substitutions on the sentence level – as opposed to the sense level – are available for 25,851 sentences.
The number of sentences that were sense-labelled with high agreement is 145,209; the total number of labelled sentences is 183,398.

5. Uses of the Resource

In this section, a few scenarios how this data could be used or extended are briefly sketched.

5.1 Lexical Substitution Task

The TWSI can be used as test data for evaluating lexical substitutions, very much like in the Semeval-2007 lexical substitution competition described by McCarthy and Navigli (2007). From this competition, 10 contexts each for 201 targets are available, with 5 annotators supplying substitutions in context. The TWSI data greatly extends the data for nouns by providing substitutions on the sentence level for over 25,000 contexts. These can be constructed from the data in the following way: subfolder `substitutions/raw_data/substitutions_per_sentence` contains for each target word the full list of substitutions per sentence. Using the sentence ID from the entry, the original sentence can be retrieved from `corpus/wiki_title_sent.txt`.
For most targets, only three annotators supplied substitutions, which makes their frequency ranking less reliable than in the lexical substitution task.

5.2 Extending the Data using Wikipedia

According to the one-sense-per-discourse hypothesis (Gale et al. 1992), repetitive uses of ambiguous words within one contextual unit (paragraph, document) are very likely to be resolved to the same meaning. Since the TWSI is drawn from Wikipedia, and the relative position of sentences in the article is retained, it is possible to automatically extend the training material by adding contexts of targets that are positioned close to a sense-labelled context for this target, using the same sense label. While this might not always succeed due to violations of the hypothesis, this might be a viable way to increase the amount of textual evidence, especially for senses with very few training examples.

6. Lexical Substitution System

The TWSI was developed to be used in a semantic search engine. A part of the system dealt with synonym expansion.
To this end, we describe a system that performs lexical substitution in context. At its core, it is realized as a supervised word sense disambiguation system (cf. Agirre and Edmonds, 2006). This system was used to create the annotations for the LRE language library for LREC 2012. A preliminary version of this system is described in more detail in (Biemann, 2010). Here, we focus on technical aspects.
As is common practice in supervised word sense disambiguation, machine learning classifiers are trained, one for each polysemous word, based on manually labelled instances. This amounts to a total of 633 classifiers for the TWSI 2.0, which imposes a challenge on memory management.
The system is implemented with the WEKA Machine Learning toolkit (Hall et al. 2009). For a given context, shallow features based on parts-of-speech, neighbouring words and content words in the vicinity are extracted. Further, to model topicality, cluster features based on word co-occurrence are supplied. For more details, please refer to Biemann (2010).

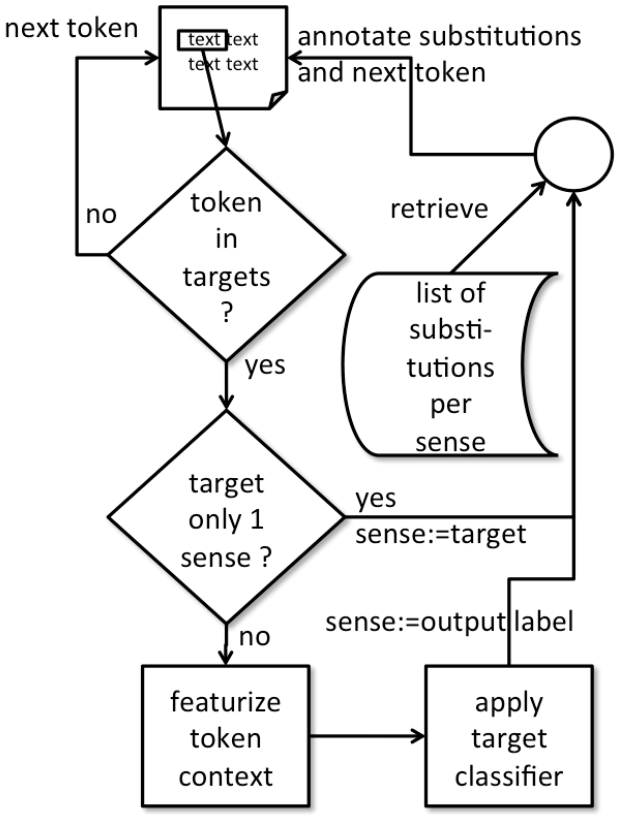

Figure 3: Flow diagram for lexical substitution system

In the training phase, instances from the `contexts` folder are characterized with features and are presented to the classifier, one polysemous target at a time, along with the sense label as found in the TWSI data. The resulting model is stored on disk for later access.

When applying the system to a given text, the text is split into tokens. Substitutions are produced for all tokens that match a list of predefined targets. For targets listed as monosemous, the substitution component simply supplies the list of substitutions for the target as provided in the `substitutions` folder. For targets that have multiple senses, the feature representation of the target instance is computed and classified by the respective classifier for the target. The classifier returns a sense label classification, together with a confidence score. For the sense label, substitutions are retrieved and supplied in context. Figure 3 depicts the flow diagram that is executed for each token of the incoming text.

Since it takes some time to load classifiers, the classifiers are loaded when needed: when a target appears for the very first time, the classifier is read from disk into memory. The next time the target appears, the classifier is already available in the memory, so its application becomes faster for further instances of the target.

In the current implementation (see final section for a download link), the machine learning method can be parameterized by the WEKA class name. The default setting, WEKA's SMO classifier (Platt, 1998), strikes a balance between model size and classification accuracy. It is possible to load the classifier models for all targets of the TWSI into less than 2GB of main memory.

The software is not restricted to the TWSI, but can be used as a supervised word sense disambiguation system for any sense-labelled data. An evaluation on a standard dataset (SemEval lexical sample task 2007; see Pradhan, 2007) showed that the system compares well against state-of-the-art methods (see Biemann (2010) for details).

To illustrate the type of annotation this software provides, Figure 4 shows the inline annotation for the first few sentences of the English Wikipedia article on "Darmstadt". For more examples, please refer to the LRE language library, where the annotation for all English written texts in the library is made available.

Darmstadt is a <target= "**city**" lemma= "city" sense= "city" confidence= "1.0" substitutions= "[town, 89] [metropolis, 50] [municipality, 40] [metropolitan area, 17] [urban area, 14] [village, 14] [urban, 13] [community, 12] [megalopolis, 12] [township, 10]"> **in the Bundesland (federal** <target= "**state**" lemma= "state" sense= "state@@3" confidence= "0.6666667" substitutions= "[government, 7] [province, 2]">) **of Hesse in Germany , located in the southern** <target= "**part**" lemma= "part" sense= "part@@1" confidence= "1.0" substitutions= "[portion, 21] [section, 21] [area, 17] [region, 15] [piece, 14] [component, 13] [segment, 11] [side, 8] [division, 6] [element, 4] [unit, 4]"> **of the Rhine Main** <target= "**Area**" lemma= "Area" sense= "area" confidence= "1.0" substitutions= "[region, 65] [zone, 24] [district, 22] [location, 21] [place, 19] [section, 17] [territory, 16] [field, 14] [part, 14] [vicinity, 14]"> **.**

The sandy <target= "**soils**" lemma= "soil" sense= "soil@@1" confidence= "1.0" substitutions= "[earth, 26] [dirt, 23] [ground, 8] [loam, 6] [land, 3] [topsoil, 2]"> **in the Darmstadt** <target= "area" lemma= "**area**" sense= "area" confidence= "1.0" substitutions= "[region, 65] [zone, 24] [district, 22] [location, 21] [place, 19] [section, 17] [territory, 16] [field, 14] [part, 14] [vicinity, 14]"> **, ill-suited for agriculture in** <target= "**times**" lemma= "time" sense= "time@@1" confidence= "0.5" substitutions= "[instance, 99] [occasion, 95] [period, 82] [moment, 60] [era, 50] [age, 24] [event, 23] [point, 22] [occurrence, 17] [duration, 16]"> **before industrial fertilisation , [2] prevented any larger** <target= "**settlement**" lemma= "settlement" sense= "settlement@@1" confidence= "1.0" substitutions= "[colony, 21] [community, 19] [village, 12] [town, 8] [hamlet, 6] [establishment, 5] [habitation, 5]"> **from developing , until the** <target= "**city**" lemma= "city" sense= "city" confidence= "1.0" substitutions= "[town, 89] [metropolis, 50] [municipality, 40] [metropolitan area, 17] [urban area, 14] [village, 14] [urban, 13] [community, 12] [megalopolis, 12] [township, 10]"> **became the** <target= "**seat**" lemma= "seat" sense= "seat@@1" confidence= "1.0" substitutions= "[position, 21] [post, 19] [place, 10] [spot, 10] [elected post, 7] [station, 5] [rank, 3] [chair position, 2] [seat of government, 2]"> **of the Landgraves of Hessen-Darmstadt in the 16th** <target= "**century**" lemma= "century" sense= "century" confidence= "1.0" substitutions= "[era, 13] [hundred, 9] [hundred year period, 9] [century period, 8] [epoch, 8] [period, 8] [age, 7] [generation, 5] [100 year period, 4] [100 years, 4] [hundred years, 4]"> **.**

Figure 4: Example output of the substitution system with inline annotation. Original text in boldface.

Each target known to the system is replaced by an inline XML-style annotation that contains the following fields:

- target: the target as spelled in the original text
- lemma: the base form of the target
- sense: the automatically classified sense
- confidence: confidence value of the classifier
- substitutions: list of weighted substitutions

The weights for the substitutions reflect the number of times this substitution was provided in the TWSI acquisition task for this sense. Annotations where the sense field does not carry sense markers (e.g. "city") supply substitutions for monosemous words. Targets with many senses get disambiguated (e.g. "state" to "state@@3)," and receive substitutions with respect to its sense. The confidence score reflects the confidence of the classifier. Both scores can be used to regularize the precision/recall trade-off for consumer processes. From the example, it becomes obvious that the TWSI resource has a considerable coverage on nouns, and provides a rich set of substitutions. While the sense classification is of high quality, not all substitutions are useful in the context. Nevertheless, this could be a stepping-stone for NLP methods, as discussed in the next section.

7. Applications of the Substitution System

Now, we discuss a few possible applications, for which a lexical substitution system could prove useful.

7.1 Semantic Search

The idea of using a substitution system for semantic search is quite straightforward: While indexing, the system is run on (suitable) documents. Besides the original targets, the substitutions are also indexed, so they can be

retrieved by a search query. In this way, the document in Figure 4 could be retrieved by a query "town with sandy topsoil", even if "town" and "topsoil" does not occur in the original text.

Substitutions can also be used for re-ranking documents returned by a keyword index, thus avoiding the necessity of processing all documents at indexing time. It remains an open question how a match between query word and substitutions should be weighted with respect to matches with the original document text. It might be advantageous, for example, to use only the most salient substitutions, to apply a threshold on the classifier confidence, and to incorporate the fact that the match came from a substitution rather than the original text into the ranking function.

7.2 Text Similarity

In applications that need to define a similarity measure between texts, such as document clustering or textual entailment, several techniques are used to go beyond a bag-of-word vector space representation (see Metzler et al. 2007). A lexical substitution system like the one described here can bridge the vocabulary gap by assigning a high similarity to text pairs where one text contains many of the substitutions of the other text. Again, it is an open question whether to use all substitutions, or only the most salient subset.

8. Conclusion

This paper presents the Turk Bootstrap Sense Inventory 2.0 (TWSI 2.0). To our knowledge, this is by far the largest electronic resource for lexical substitution in existence, and it contains the largest collection of sense annotations in the style of the lexical sample task (i.e. a high number of occurrences for a limited set of words). In previous research using only that part of the data that was released as TWSI 1.0, it was shown that this resource enables lexical substitution systems with substitution acceptance rates well over 90% (Biemann, 2010). It is therefore a building block for a wide range of NLP tasks.

Further, we described a lexical substitution system trained on this resource. It uses supervised word sense disambiguation techniques to supply substitutions in context by applying a machine learning classifier for each target and annotating the target with the substitutions corresponding to the automatically identified sense. Possible uses and applications of both resource and substitution system were laid out briefly.

The resource is available for download under a Creative Commons Share-Alike license at *http://www.ukp.tu-darmstadt.de/data/twsi-lexical-substitutions/*. The lexical substitution system described in Section 5 is available for download as an open-source Java project at *http://www.ukp.tu-darmstadt.de/software/twsi-sense-substituter/* under the GPL license.

9. Acknowledgements

The creation of the data has been fully funded by the Microsoft Corporation. The lexical substitution system software was supported by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz"" (LOEWE) as part of the research centre "Digital Humanities".

10. References

Agirre, E., Edmonds, P. (2006): *Word Sense Disambiguation: Algorithms and Applications* (Text, Speech and Language Technology), Springer-Verlag New York

Biemann, C. and Nygaard, V. (2010): Crowdsourcing WordNet. *Proceedings of the 5th Global WordNet Conference*, Mumbai, India.

Biemann, C. (2010): Co-occurrence Cluster Features for Lexical Substitutions in Context. *Proceedings of the 5th Workshop on TextGraphs in conjunction with ACL 2010*, Uppsala, Sweden

Gale, W.A., Church, K.W., and Yarowsky, D. (1992): One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pp. 233-237

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009); The WEKA Data Mining Software: An Update; *SIGKDD Explorations, Volume 11, Issue 1*

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L, and Weischedel, R. (2006): OntoNotes: The 90% solution. In *Proceedings of HLT-NAACL* 2006, pages 57–60.

McCarthy, D. and Navigli, R. (2007): SemEval-2007 Task 10: English Lexical Substitution Task. *Proc. of Semeval-2007 Workshop (SEMEVAL),* in the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic

Metzler, D., Dumais, S., and Meek, C. (2007). Similarity measures for short segments of text. *Proceedings of the 29th European conference on IR research (ECIR-07)* pp. 16-27, Springer

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., Miller, K. (1990): WordNet: An online lexical database. *Int. J. Lexicograph*. 3, 4, pp. 235–244.

Platt, J. (1998): Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*

Pradhan, S., Loper, E., Dligach, D. and Palmer, M. (2007) SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007),* pp. 87–92, Prague, Czech Republic

Sinha, R., McCarthy, D. and Mihalcea, R. (2009): SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009),* Boulder, Colorado