

# Word Sense Distribution in a Web Corpus

Ping Chen, David Brown, Andrew Tran, Noble Ozoka, Rafael Ortiz  
Department of Computer and Mathematics Sciences  
University of Houston-Downtown  
1 Main St., Houston, TX, USA

## Abstract

World Wide Web has become an important knowledge source for many research fields, and quality of Web-acquired knowledge has direct impact on their performance. While evaluation of the vast amount of Web resources is out of question, in this paper we examined thousands of sentences containing twelve pre-selected words and produced several quality measures including sentence coherence and sense distribution information. Our goal is to provide some insight to several Computational Linguistics areas that acquire knowledge from the Web.

**KEYWORDS:** Word sense distribution, Web corpus acquisition and quality analysis, Sense annotation, Computational Linguistics

## 1 Introduction

Many Computational Linguistics (CL) tasks are knowledge-intensive by nature. Take word sense disambiguation (WSD) as an example, a practical WSD system needs to contain disambiguation-capable knowledge about a broad range of words and their various usage. Such a requirement is not easily satisfied since a natural language usually contains thousands of words, and some words can have dozens of senses. For example, the Oxford English Dictionary has approximately 301,100 main entries [14], and the average polysemy of the WordNet inventory is 6.18 [5]. Moreover, a natural language is not a static phenomenon. New usage of existing words emerges, which creates new senses. New words are

created, and some words may “die” over time. It is estimated that every year around 2,500 new words appear in English [6]. Such dynamics requires constant and timely maintenance and updating of existing knowledge bases.

In recent years many research projects have adopted the Web as one major knowledge source, e.g., KnowItAll, a domain-independent knowledge acquisition system [4], verb information collection [3], extraction of IS-A relations from Web text [11], N-gram collection [2]. Undoubtedly, with billions of diverse documents readily accessible and millions of web pages created or updated daily, the Web provides a broad and dynamic coverage of many languages. However, such diversity also inevitably results in resources of varying quality, from totally nonsense pages (e.g., spam web pages) to high-quality articles. Effectively assessing coverage and quality of Web documents is critically important to Web-based knowledge acquisition methods, since many CL applications have become increasingly dependent on such methods.

Considering the vast amount of Web documents, automatic assessment tools are certainly desirable. However, automatic assessment has proved to be rather difficult and cannot produce reliable results [12]. In this paper, we describe one experiment to manually evaluate the Web documents collected through a Web search engine. Comparing with general Web crawling, keyword-based search provides targeted text, and can also filter out many spam or low-quality Web pages. While any manual evaluation is inevitably small-scale, results generated from educated human beings are still considered far more

trustworthy. To minimize any bias that may result from analysis of a small amount of text, manual assessments require careful selection of evaluation metrics and target text. In this paper we will discuss one initial experiment to analyze all sentences obtained through a search engine that contain instances of twelve English words preselected according to several diverse criteria. Although it is only a small set of words, for each word we analyzed a substantial number of its instances acquired through a search engine, so we believe that in this aspect our analysis of each word is exhaustive and complete.

This paper is organized as follows. Related work is given in Section 2. We discuss the experiment setup in Section 3, and experiment results in section 4. We conclude in section 5.

## 2 Related work

Due to the importance of high-quality annotated corpus in CL, there have been many efforts to produce large general-purpose corpus with different types of annotations. The Penn Treebank Project annotates naturally-occurring text with syntactic and semantic information and part-of-speech tags. Treebank Project started in 1989, and one release Treebank 2 contains over 1.6 million words of hand-parsed material, an additional one million words tagged for part-of-speech, and also the first fully parsed version of the Brown Corpus [8]. The Proposition Bank project adds a layer of predicate-argument information, or semantic role labels, to the syntactic structures of the Treebank [10]. Both Treebank and Proposition Bank take a hybrid annotation approach that starts with automatic tools and follows by manual corrections. The closest project to our work may be Sencor. Sencor includes text extracted from the Brown corpus, and words are syntactically and semantically tagged with Brill’s part of speech tagger and senses from WordNet[13].

## 3 Experiment

The goal of our work is to find out the quality of the Web documents returned by a search engine and sense distributions based on a given set of words. Sense distribution information is very useful in many

CL fields. For example, in word sense disambiguation, most frequent sense baseline system (simply choosing most frequent sense as the correct sense) often performs better than many sophisticated WSD systems. When constructing a dictionary, lexicographers spend an enormous amount of time collecting representative word instances of diverse usage, ensuring that each and every sense is represented sufficiently and accurately. If the Web, a readily available resource to researchers and lexicographers, can provide high quality resources, Web resources will be adopted with more confidence in more applications.

### 3.1 Word Selection

The first step of this experiment is to choose a set of words to collect their instances. The words were chosen with the following characteristics in mind:

1. A characteristic distribution of senses across parts of speech (noun, verb, adverb, and adjective)
  - (a) containing one or two parts of speech
  - (b) containing all four parts of speech
2. A characteristic number of senses
  - (a) containing only 1 or 2 senses
  - (b) containing many senses (e.g., dozens of senses)
3. A range of frequency: some are rare and some very common words, some are common in one part of speech but rare in others

The words themselves were chosen using four English word repositories, American Heritage most frequent word list (3 words are chosen: back, long, put), British National Corpus (BNC) frequently spoken word list (3 words are chosen: center, mind, case), BNC frequency 03 list (3 words are chosen: author, cart, core), and BNC frequency 04 list (3 words are chosen: dissolve, sequence, toast). These words were selected after looking through many words and examining their sense distributions, popularity, and overall number of senses. Only twelve were chosen because of the enormous amount of time and labor needed to manually annotate the collected sentences and analyze the results. By choosing words with

characteristic diverse distributions spreading across different parts of speech, frequencies, and number of senses, hopefully what we learn from instances of these twelve words applies to a wider vocabulary in English as well.

### 3.2 Word instance crawling

The preselected twelve words were submitted to MSN LiveSearch, and 1000 HTML documents and 1000 text documents were returned. The collection of documents (all 2000 results for all twelve words) took approximately 3 hours, then sentences containing the preselected words were extracted and compiled into an individual file.

### 3.3 Annotation

The annotation of the sentences containing preselected words is done by two annotators for quality control. Each annotator was given an identical copy of the sentences and an identical set of tagging instructions and individually tagged all of these sentences. The annotators complete annotation of sentences for one word (anywhere from 1000 to 3000 sentences) on an average of three to five days. Here are the annotation tags:

1. Part Of Speech: can be assigned as noun, verb, adverb, or adjective.
2. Sense ID: Corresponds to the sense ID's provided by WordNet.
3. Coherence Level: There are 4 possible values assigned by the annotator,
  - (a) 0 - The sentence has no problems at all, i.e. "You are now a sense annotator."
  - (b) 1 - The sentence has structural or syntactical issues, i.e. "put down, but Greg didn't stop there whenever."
  - (c) 2 - The sentence is syntactically correct but contains semantic problems, i.e. "Green ideas sleep furiously."
  - (d) 3 - The sentence has no grammatical, syntactical, or semantic problems, but is factually incorrect, i.e. "Many human beings live on Mars in 2010."

## 4 Annotation results

Totally 14,366 sentences were collected. The annotators are given the WordNet repository in order to determine part of speech and sense of each target word. Once these are determined, the annotators give the sentence a subjective coherence score. This coherence score reflects the linguistic quality of these sentences. If the annotations on a sentence from two annotators are different, a third annotator will try to break the tie. If he fails, this sentence will not be included in the final experiment results.

### 4.1 Quality of the Web corpus

After annotation is finished, 11,160 sentences are recorded after inter-annotator examination. The results are shown in Table 1. 70.86% of these sentences are both syntactically and semantically correct. 28.80% of sentences contain syntactic problems. But semantically this Web corpus is of very high quality, only 0.27% of sentences violate common sense knowledge, and 0.06% are factually wrong. These results indicate that text returned by search engines are of high quality in general. While some sentences contain syntax errors, semantic or factual errors are quite rare.

### 4.2 Sense distribution results

After examination of the third annotator, results from 11,232 sentences were recorded as shown in Table 2. For comparison purpose, we also included the number of tagged instances from WordNet in the table. For example, as a noun, sense 1 of "author" appears 990 times in our collected sentences, and 38 times from tagged text in WordNet. The number after part of speech shows the number of senses of a word taking this part of speech. For example, "noun:2" means that as a noun this word has 2 senses in WordNet repository. In the table S1, ..., S9 indicate sense1, ..., sense9 respectively based on WordNet repository. For the most frequent senses, we found that there are 3 disagreements out of 28 different parts of speech of the 10 target words. Among the 146 senses of the 10 target words, our collection contains instances of 116 senses, while WordNet contains instances of 104 senses. Generally sense distribution

Word	POS	C=0	1	2	3
author	noun	635	433	2	4
	verb	77	33	0	0
back	noun	33	21	1	0
	verb	89	33	0	0
	adj.	36	14	1	0
	adv.	547	232	8	2
cart	noun	554	165	2	0
case	noun	778	334	3	0
center	noun	298	136	3	0
	verb	3	0	0	0
core	noun	582	136	1	0
dissolve	noun	15	3	0	0
	verb	784	309	1	0
long	verb	4	1	0	0
	adj.	760	244	1	1
	adv.	9	1	0	0
mind	noun	724	272	2	0
	verb	21	8	0	0
put	verb	672	254	2	0
sequence	noun	733	352	3	0
toast	noun	385	148	0	0
	verb	169	86	0	0
Percentage(%)		70.86	28.80	0.27	0.06

Table 1: Quality of collected sentences containing the 12 target words. C=0, 1, 2, 3 indicate the coherence level of sentences.

of our Web corpus is quite similar with WordNet, while a Web corpus may provide more diverse senses and far more instances of each individual word.

## 5 Conclusion

As the Web becomes an increasingly important knowledge source in many Computational Linguistics fields, assessment of Web resource quality deserves immediate investigation. In this paper we collected a large number of sentences through a Web search engine containing 12 preselected words, and manually annotate these sentences. Our main findings are: (1) Web text may contain syntax errors, but semantic or factual errors are rare; (2) Most frequent sense information from a Web corpus is similar as WordNet; (3) Web can provide a more diverse sense coverage and

far more instances containing a specific word. We will make our annotated Web corpus freely available online for CL research community.

## Acknowledgments

This work is funded by National Science Foundation grant CNS 0851984 and Department of Homeland Security grant 2009-ST-061-C10001.

## References

- [1] E. Agirre, Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*, Springer. 2006.
- [2] S. Bergsma, D. Lin, R. Goebel. *Web-Scale N-gram Models for Lexical Disambiguation*, Proceedings of IJCAI, 2009.
- [3] T. Chklovski, P. Pantel. *VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations*. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*. pp. 33-40. Barcelona, Spain, 2004.
- [4] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. *Unsupervised named-entity extraction from the web: An experimental study*. *Artificial Intelligence*, 165(1):91-134. 2005
- [5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT press. 1998.
- [6] K. Kister. *Dictionaries defined*. *Library Journal*, v117 n11 p43-46. 1992.
- [7] LDC Catalog. <http://www ldc.upenn.edu/>
- [8] M. Marcus, M. A. Marcinkiewicz, and B. Santorini. *Building a large annotated corpus of English: the penn treebank*. *Comput. Linguist.* 19, 2 (Jun. 1993), 313-330.
- [9] D. McCarthy, Rob Koeling, Julie Weeds, John A. Carroll. *Finding Predominant Word Senses In Untagged Text*, ACL, 2004.

Word	POS	S1	S2	S3	S4	S5	S6	S7	S8	S9
author	noun:2	990/38	84/6							
	verb:1	110/0								
back	noun:9	28/53	14/12	3/4	9/0	0/0	0/0	0/0	0/0	0/0
	verb:10	81/7	9/6	2/4	4/4	3/2	3/0	2/0	0/0	17/0
	adj.:3	42/15	6/1	5/0						
	adv.:6	260/92	77/36	236/24	138/15	18/14	75/1			
cart	noun:2	272/5	450/1							
	verb:2	0/0	0/0							
case	noun:18	489/705	209/703	55/237	90/106	97/98	64/84	66/48	51/45	0/34
	verb:2	0/0	0/0							
center	noun:18	71/56	26/10	261/6	31/5	11/3	23/3	0/3	0/2	19/1
	verb:3	1/12	0/4	1/0						
	adj.:2	0/0	0/0							
core	noun:9	187/5	93/5	1/2	440/2	1/1	0/0	0/0	2/0	5/0
	verb:1	0/0								
dissolve	noun:1	16/4								
	verb:11	452/17	50/12	101/4	121/2	3/1	0/0	3/0	3/0	92/0
long	verb:1	5/6								
	adj.:9	863/118	129/107	6/2	3/0	3/0	1/0	9/0	0/0	0/0
	adv.:2	10/37	0/0							
mind	noun:7	643/121	94/13	116/10	23/3	78/1	38/0	8/0		
	verb:6	29/15	0/5	0/2	0/0	0/0	0/0			
put	noun:1	0/0								
	verb:9	579/691	140/253	64/125	86/81	37/15	4/5	19/4	0/3	1/3
sequence	noun:5	998/13	19/3	73/3	0/3	0/0				
	verb:2	0/1	0/0							
toast	noun:4	350/5	38/0	96/0	50/0					
	verb:2	104/5	153/1							

Table 2: Sense distribution of the 12 target words

- [10] M. Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71-106, 2005.
- [11] P. Pantel, D. Ravichandran, E. Hovy. Towards Terascale Knowledge Acquisition. In *Proceedings of Conference on Computational Linguistics (COLING-04)*. pp. 771-777. Geneva, Switzerland, 2004.
- [12] I. Ryu, Diana McCarthy, Rob Koeling. Gloss-Based Semantic Similarity Metrics for Predominant Sense Acquisition, *Proceedings of the Third International Joint Conference on Natural Language Processing*, 2008.
- [13] SemCor. [www.cse.unt.edu/rada/downloads.html](http://www.cse.unt.edu/rada/downloads.html)
- [14] C. Soanes, Angus Stevenson, editors. *Oxford Dictionary of English*. Oxford University Press. 2003.