Expectations of Word Sense in Parallel Corpora

Xuchen Yao, Benjamin Van Durme and **Chris Callison-Burch**
Center for Language and Speech Processing, and HLTCOE
Johns Hopkins University

Abstract

Given a parallel corpus, if two distinct words in language A, a_1 and a_2, are aligned to the same word b_1 in language B, then this might signal that b_1 is polysemous, *or* it might signal a_1 and a_2 are synonyms. Both assumptions with successful work have been put forward in the literature. We investigate these assumptions, along with other questions of word sense, by looking at sampled parallel sentences containing tokens of the same type in English, asking how often they mean the same thing when they are: 1. aligned to the *same* foreign type; and 2. aligned to *different* foreign types. Results for French-English and Chinese-English parallel corpora show similar behavior: Synonymy is only very weakly the more prevalent scenario, where both cases regularly occur.

1 Introduction

Parallel corpora have been used for both paraphrase induction and word sense disambiguation (WSD). Usually one of the following two assumptions is made for these tasks:

1. **Polysemy** If two different words in language A are aligned to the same word in language B, then the word in language B is polysemous.

2. **Synonymy** If two different words in language A are aligned to the same word in language B, then the two words in A are synonyms, and thus is *not* evidence of polysemy in B.

Despite the alternate nature of these assumptions, both have associated articles in which a researcher claimed success. Under the polysemy assumption, Gale et al. (1992) used French translations as English sense indicators in the task of WSD. For instance, for the English word *duty*, the French translation *droit* was taken to signal its *tax* sense and *devoir* to signal its *obligation* sense. These French words were used as labels for different English senses. Similarly, in a cross-lingual WSD setting,[1] Lefever et al. (2011) treated each English-foreign alignment as a so-called *ParaSense*, using it as a proxy for human labeled training data.

Under the synonymy assumption, Diab and Resnik (2002) did word sense tagging by grouping together all English words that are translated into the same French word and by further enforcing that the majority sense for these English words was projected as the sense for the French word. Bannard and Callison-Burch (2005) applied the idea that French phrases aligned to the same English phrase are paraphrases in a system that induces paraphrases by pivoting through aligned foreign phrases.

Based on this, and other successful prior work, it seems neither of the assumptions must hold universally. Therefore we investigate how often we might *expect* one or the other to dominate: we sample polysemous words from wide-domain {French,Chinese}-English corpora, and use Amazon's Mechanical Turk (MTurk) to annotate word sense on the English side. We calculate empirical probabilities based on counting over the competing polysemous and synonymous scenario labels.

A key factor deciding the validity of our conclusion is the reliability of the annotations derived via MTurk. Thus our first step is to evaluate the ability of Turkers to perform WSD. After verifying this

[1] E.g., given a sentence *"... more power, more duty ..."*, the task asks to give a French translation of *duty*, which should be *devior*, after first recognizing the underlying *obligation* sense.

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 621–625,
Montréal, Canada, June 3-8, 2012.

as a reasonable process for acquiring large amounts of WSD labeled data, we go on to frame the experimental design, giving final results in Sec. 4.

2 Turker Reliability

While Amazon's Mechanical Turk (MTurk) has been been considered in the past for constructing lexical semantic resources (e.g., (Snow et al., 2008; Akkaya et al., 2010; Parent and Eskenazi, 2010; Rumshisky, 2011)), word sense annotation is sensitive to subjectivity and usually achieves low agreement rate even among experts. Thus we first asked Turkers to re-annotate a sample of existing gold-standard data. With an eye towards costs saving, we also considered how many Turkers would be needed per item to produce results of sufficient quality.

Turkers were presented sentences from the test portion of the word sense induction task of SemEval-2007 (Agirre and Soroa, 2007), covering 2,559 instances of 35 nouns, expert-annotated with OntoNotes (Hovy et al., 2006) senses. Two versions of the task were designed:

1. **compare**: given the same word in different sentences, tell whether their meaning is THE SAME, ALMOST THE SAME, UNLIKELY THE SAME or DIFFERENT, where the results were collapsed post-hoc into a binary same/different categorization;

2. **sense map**: map the meaning of a given word in a sentential context to its proper OntoNotes definition.

For both tasks, $2,599$ examples were presented.

We measure inter-coder agreement using Krippendorff's Alpha (Krippendorff, 2004; Artstein and Poesio, 2008), where $\alpha \geq 0.8$ is considered to be reliable and $0.667 \leq \alpha < 0.8$ allows for tentative conclusions. Two points emerge from Table 1: there were greater agreement rates for sense map than compare, and 3 Turkers were sufficient.

	α-Turker	α-maj.	maj.-agr.
compare$_5$	0.47	0.66	0.87
compare$_3$	0.44	0.52	0.83
sense map$_5$	0.79	0.93	0.95
sense map$_3$	0.75	0.87	0.91

Table 1: MTurk result on testing Turker reliability. Krippendorff's Alpha is used to measure agreement. α-Turker: how Turkers agree among themselves, α-maj.: how the majority agrees with true value, maj.-agr.: agreement between the majority vote and true value. α-maj. indicates the confidence level about the maj.-agr. value. Subscripts denote either 5 Turkers, or 3 randomly selected of the 5.

3 Experiment Design

Data Selection We used two parallel corpora: the French-English 10^9 corpus (Callison-Burch et al., 2009) and the GALE Chinese-English corpus.

For each corpus we selected 50 words, w, at random from OntoNotes,[2] constrained such that w: had more than one sense; had a frequency $\geq 1,000$; and was not a top 10% most frequent words.

Next we sampled 100 instances (aligned English-foreign sentence pairs) for each word based on the following constraints: the aligned foreign word, f, had a frequency ≥ 20 in the foreign corpus; f had a non-trivial alignment probability.[3] We sampled proportionally to the distribution of the aligned foreign words, ensuring that at least 5 instances from each foreign translation are sampled.[4]

For each corpus, this results in 100 instances for each of 50 words, totaling 5,000 instances. We used 3 Turkers per instance for sense annotation, under the sense map task. We note that the set of 50 randomly selected English words from the Chinese-English corpus were entirely distinct from the 50 selected words from the French-English corpus.

Probability Estimation Suppose e_1 and e_2 are two tokens of the same English word type e. $s(e_1)$ is a function that returns the sense of e_1, $a(e_1)$ is a function that returns the aligned word of e_1. Let $c()$ be our count function, where: $c(e, f)$ returns the

[2]OntoNotes was used as the sense inventory over alternatives, owing to its coarse-grained sense definitions.

[3]Defined as f having index $i < k$ when foreign words are ranked by most probable given e, where k is the minimum value such that $\sum_i^k p(f_i \mid e) > 0.8$. E.g., if we have decreasing probabilities $p(droit \mid duty) = 0.6$, $p(devoir \mid duty) = 0.25$, $p(le \mid duty) = 0.03$, ... then only consider *droit* and *devoir*. This ruled out many noisy alignments.

[4]Thus, the instances of *droit* compared to that of *devoir* would be $0.6/0.25$.

number of times English word e is aligned to foreign word f; $c(e^s, f)$ returns the number of times English word e has sense s (tagged by Turkers), when aligned to foreign word f; $c(e)$ is the total number of tokens of English word e; and $c(e^s)$ is the number of tokens of e with sense s.

We estimate from labeled data the probability of three scenarios, with scenario 1 as our primary concern: when two English words of the same polysemous type are aligned to *different* foreign word types, what is the chance that they have the same sense? Given the tokens e_1 and e_2, we calculate P1 as follows:

$$\begin{aligned} P1_e &= P(s(e_1) = s(e_2) \mid a(e_1) \neq a(e_2)) \\ &\approx \frac{\sum_s c^2(e^s) - \sum_{s,f} c^2(e^s, f)}{c^2(e) - \sum_f c^2(e, f)} \end{aligned}$$

P1 says that given two words of the same type (e_1 and e_2) that are *not* aligned to the same foreign word type ($a(e_1) \neq a(e_2)$), what is the probability that they have the same sense ($s(e_1) = s(e_2)$). We approach this estimation combinatorially. For instance, the number of ways to choose two words of the same type is $\binom{2}{c(e)} \approx \frac{1}{2}c^2(e)$ when $c(e)$ is large.

A large value of P1 would be in support of **Synonymy**, as the two foreign aligned words of distinct type would have the same meaning.

Scenario 2 asks: given two English words of the same polysemous type and aligned to the *same* words ($a(e_1) = a(e_2)$), what is the probability that they have the same sense ($s(e_1) = s(e_2)$)?

$$\begin{aligned} P2_e &= P(s(e_1) = s(e_2) \mid a(e_1) = a(e_2)) \\ &\approx \frac{\sum_{s,f} c^2(e^s, f)}{\sum_f c^2(e, f)} \end{aligned}$$

Finally, what is the probability of two tokens of the same polysemous type agreeing when alignment information is not known (e.g., without a parallel corpus)?

$$P3_e = P(s(e_1) = s(e_2)) \approx \frac{\sum_s c^2(e^s)}{c^2(e)}$$

All the above equations are given per English word type e. In later sections we report the average values over multiple word types and their counts.

4 Results

Turker Experiments To minimize errors from Turkers, for every HIT we inserted one control sentence taken from the example sentences of OntoNotes. Turker results with either extremely low finishing time (<10s), or average accuracy on control sentences lower than accuracy by chance, were rejected. On average Turkers took 185 seconds to map 10 sentences in a HIT to their OntoNotes definition, receiving $0.10 per HIT. The total time for annotating 5000 sentences was 22 hours.

Turkers had no knowledge about alignments: we hid the aligned French/Chinese sentences from them and these sentences were later processed to compute P1/2/3 values. Two foreign tokens aligned with the same source type correspond to two senses of the same type. To give an estimate of alignment errors, we manually examined 1/10 of all 5000 sampled Chinese-English alignments at random and found only 3 of them were wrong: all due to that English content words were aligned to common Chinese function words. This error rate is much lower than that typically reported by alignment tools. The main reason is explained in footnote 3: foreign words with trivial alignment probability were removed before calculating P1/2/3 values. Thus we believe the alignment was reliable.

Probability Estimation Table 2 gives the distribution of senses and word types in the sampled words. Take the second numeric column of French-English as an example: out of 50 words randomly sampled, 9 have 2 distinct sense definitions in OntoNotes. However, 17 of 50 unique word types had exactly 2 distinct senses annotated, out of the 100 examples of a given word type: 17 words had 2 distinct senses *observed*. Of the 9 words with 2 official senses, on average 1.9 of those senses were observed.

Table 3 and Figures 1 and 2 shows the result for P1, P2 and P3 using the {French,Chinese}-English corpora, calculated based on the majority vote of three Turkers. High P2 values suggests that for two tokens of the same type, aligning to the same foreign type is a reasonable indicator of having the same meaning. When working with open domain corpora, without foreign alignments, the probability of two English words of the same type having

	French-English									Chinese-English									
#senses in OntoNotes	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	18
#types in OntoNotes	0	9	7	6	8	6	2	8	4	0	10	6	11	3	8	6	4	1	1
#types observed	2	17	9	4	7	7	4	0	0	3	19	9	12	5	2	0	0	0	0
avg #senses observed	0	1.9	2.1	3.2	3.8	4.7	6.5	4.9	5.8	0	1.9	2.2	2.9	2.7	4.4	3.8	3.8	3.0	5.0

Table 2: Statistics for words sampled from parallel corpora. Average #senses observed over all words: 2.6 (French-English), and 2.4 (Chinese-English). The sampled word *keep* has 18 senses in OntoNotes, with 5 observed.

	P1	P2	P3	Alpha
French-English	51.2%	66.7%	59.2%	0.70
Chinese-English	59.6%	78.7%	66.7%	0.68

Table 3: Expectations of word sense in parallel corpora. Alpha measures how Turkers agreed with themselves.

identical meaning is estimated here to be roughly 59-67% (59.2% (French), 66.7% (Chinese)). This accords with results from WSD evaluations, where the first-sense heuristic is roughly 75-80% accurate (e.g., 80.9% in SemEval'07 (Brody and Lapata, 2009)). Minor algebra translates this into an expected P3 value in a range from $56\% - 62.5\%$, up to $64\% - 68\%$, which captures our estimates.[5]

Finally for our motivating scenario: values for P1 are barely higher than 50%, suggesting that **Synonymy** more regularly holds, but not conclusively. We expect in narrower domains, where words have less number of senses, this is more noticeable. As suggested by Fig.s 1 and 2, less polysemous words tend to have higher P values.

5 Conclusion

Curious as to the distinct threads of prior work based on alternate assumptions of word sense and parallel corpora, we derived empirical expectations on the shared meaning of tokens of the same type appearing in the same corpus. Our results suggest neither the assumption of Polysemy nor Synonymy holds significantly more often than the other, at least for individual words (as opposed to phrases) and for the open domain corpora used here. Further, we provide an independent data point that supports earlier findings as to the expected accuracy of the first sense heuristic in word sense disambiguation.

[5]Assuming worst case: no two tokens that are not the first sense ever match, and best case: any two tokens not the first sense always match, then assuming first-sense accuracy of 0.8 gives a range on P3 of: $(0.8^2, 0.8^2 + 0.2^2) = (0.64, 0.68)$.

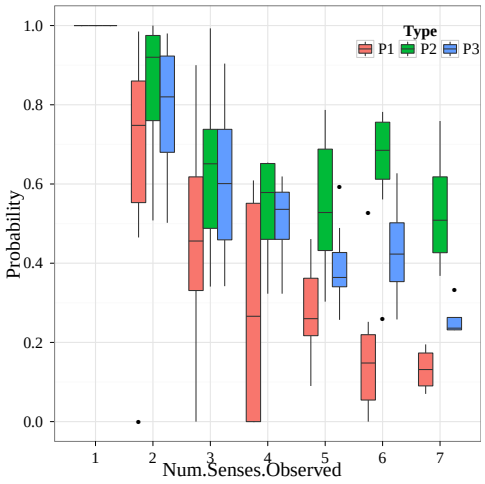

Figure 1: French-English values, by number of senses.

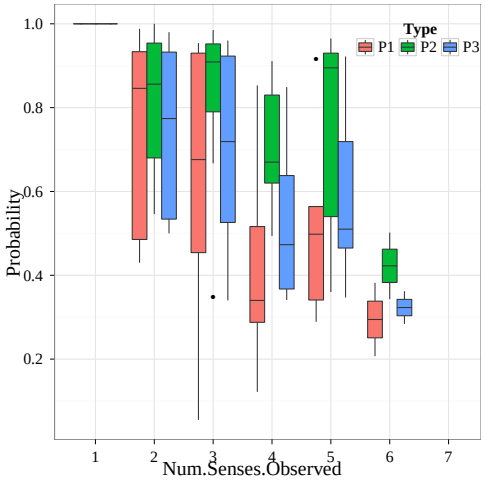

Figure 2: Chinese-English values, by number of senses.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 Task 02: Evaluating Word Sense Induction And Discrimination Systems. In *Proc. SemEval '07*.

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proc. NAACL Workshop on CSLDAMT*.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4).

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. ACL*.

Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proc. EACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings Of The 2009 Workshop On Statistical Machine Translation. In *Proc. StatMT*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proc. ACL*.

W.A. Gale, K.W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proc. TMI*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proc. NAACL-Short*.

Klaus H. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proc. ACL*.

Gabriel Parent and Maxine Eskenazi. 2010. Clustering dictionary definitions using amazon mechanical turk. In *Proc. NAACL Workshop on CSLDAMT*.

Anna Rumshisky. 2011. Crowdsourcing word sense definition. In *Proc. LAW V*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP*.