Using semi-experts to derive judgments on word sense alignment: a pilot study

Soojeong Eom, Markus Dickinson, Graham Katz

Georgetown University, Indiana University, Georgetown University
se48@georgetown.edu, md7@indiana.edu, egk7@georgetown.edu

Abstract

The overall goal of this project is to evaluate the performance of word sense alignment (WSA) systems, focusing on obtaining examples appropriate to language learners. Building a gold standard dataset based on human expert judgments is costly in time and labor, and thus we gauge the utility of using semi-experts in performing the annotation. In an online survey, we present a sense of a target word from one dictionary with senses from the other dictionary, asking for judgments of relatedness. We note the difficulty of agreement, yet the utility in using such results to evaluate WSA work. We find that one's treatment of *related* senses heavily impacts the results for WSA.

Keywords: word sense alignment, crowdsourcing, evaluation, WordNet, COBUILD

1. Introduction

As is well-known, different word sense inventories contain non-trivial mappings between them. As one example, (Palmer et al., 2000) discuss the various problems in aligning the senses of *shake* in Hector (Atkins, 1993) with those in WordNet (Fellbaum, 1998), such as the TREMBLE and MOVE distinctions in Hector being conflated in WordNet. Such differences arise from the different purposes of sense inventories, e.g., for lexicography, computational disambiguation systems, or relevance to language learners. Each inventory also has its own particular design, with some including hierarchical information, some having illustrative examples for senses, some based on thesaurus information, and so forth.

Partly owing to these differences, there is a need to align senses between inventories for a number of applications, such as, e.g., building large-scale lexical databases for machine translation which combine sources of information (Knight and Luk, 1994); comparing the performance of different natural language processing (NLP) systems characterizing lexical semantics (Nastase and Szpakowicz, 2001); or reducing the granularity of an inventory for NLP (Navigli, 2006). This last case is particularly important, as aligning senses to coarser-grained ones can lead to high-performance word sense disambiguation (WSD) systems (Navigli, 2009; Navigli et al., 2007). Indeed, increasing the scale of sense inventories is an ongoing task, especially for newer resources such as Wikipedia and Wiktionary (see, e.g., (Meyer and Gurevych, 2011) and references therein).

Our starting point comes from a different perspective yet. We intend to map word senses between two sense inventories, to link up state-of-the-art automatic WSD systems employing WordNet with a sense resource containing examples more appropriate to display for language learners, namely a resource like the COBUILD dictionary (Sinclair, 2006). COBUILD is a dictionary specifically designed for learners of English, and, although proprietary, it is used by learners around the world.[1] We are developing an online system to provide vocabulary assistance to learners of English, allowing them to click on unfamiliar words and see examples relevant to that usage (cf. (Nerbonne and Smit, 1996; Heilman et al., 2006)). To do this, we need to develop a system which automatically maps from WordNet (output from an automatic WSD system) to something like COBUILD (for displaying examples).

This is a much different goal than much of the other alignment work, in that we do not need to expand a resource and make it bigger, but instead map consistently from one to the other. Further, this mapping between the inventories needs to be reliable, as false mappings can lead learners astray—i.e., are worse than no mapping at all.

Thus, in this paper, we investigate the upper bound on automatic word sense alignment (WSA) accuracy by testing human accuracy. Additionally, we need evaluation data for WSA; as far as we know, no such database exists for the types of inventories we are interested in, though other databases exist, predominantly for ones linking WordNet and Wikipedia or Wiktionary (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011; Wolf and Gurevych, 2010; Fernando and Stevenson, 2010; Toral et al., 2009). We want to develop a gold standard of alignments between the inventories of interest, and we provide a methodology here which allows us to do so. Specifically, we explore pooling judgments from semi-experts—i.e., linguistic students and faculty. While this has the potential to speed up annotation efforts, we employ this methodology mainly to get a better grasp on how much people with linguistic knowledge can agree on meaning similarity.

The main contributions of this paper are to: 1) explore alignments between two resources which have not been robustly investigated (WordNet and COBUILD); 2) add to the line of work allowing for gradable judgments for lexical semantics (Erk and McCarthy, 2009; Erk et al., 2009); and 3) test the effectiveness of using semi-experts to provide similarity judgments.

[1] e.g., `http://endic.naver.com`

2. Related work

Our work is intended to be useful for word sense alignment (WSA), so we first review some of that literature, focusing on which inventories are aligned and what evaluation data is used (see also (Meyer and Gurevych, 2011) for a good overview of WSA work). To start with, (Ide and Véronis, 1990) combine dictionaries (the Collins English Dictionary (CED) and Oxford Advanced Learner's Dictionary (OALD)) to create a comprehensive knowledge base. (Knight and Luk, 1994) construct a large-scale knowledge base for machine translation by merging existing resources (WordNet and the Longman Dictionary Of Contemporary English (LDOCE)). (Kwong, 1998) is similar, but incorporates Roget's Thesaurus and further organizes the resource. In some of these cases and others (Ruiz-Casado et al., 2005; Nastase and Szpakowicz, 2001), the evaluation seems to have been done by manual evaluation by a single annotator (or was not specified). Much recent work has focused on having multiple annotators perform this task, to better gauge levels of agreement (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011; Navigli and Ponzetto, 2010; Navigli, 2006).

To take one recent example, (Meyer and Gurevych, 2011) align between WordNet and Wiktionary, following a long line of work on mapping between WordNet and either Wikipedia or Wiktionary (e.g., (Navigli and Ponzetto, 2010; Ruiz-Casado et al., 2005; Toral et al., 2009)). Interestingly, they ask for judgments on 2,423 sense pairs about whether the senses have the same meaning or a different meaning. Our approach, on the other hand, allows for some graded notion of meaning, i.e., a *related meaning* category. This is in line with what (Meyer and Gurevych, 2011) note in their error analysis about mis-alignment, where one often wants to link senses with related meanings: "Future work could distinguish between sense alignments sharing the same meaning and sharing a highly related meaning."

Turning from word sense alignment, similar in spirit to our work is work comparing judgments on word senses across different contexts. Prominently, (Erk and McCarthy, 2009; Erk et al., 2009) explore graded word sense judgments, i.e., allowing annotators to select the degree of similarity for a word sense on a given task, not just restricting the task to selecting a single sense.

(Erk et al., 2009), for instance, performed two experiments. In the first, **WSsim** (Word Sense Similarity), they asked annotators to read sentences and, for *every* WordNet sense, assign a similarity score, between 1 (*completely different*) and 5 (*identical*). This allows annotators to grade all senses, instead of making a binary choice for each sense or even selecting a single sense. In the second experiment, **Usim** (Usage Similarity), annotators were given pairs of sentences and ask to rank how similar in meaning the two usages of a given word were (using the same 5-point scale). What they found was that "[t]he annotators made use of the full spectrum of ratings."

Our experiment is similar in spirit and design with the **USim**, in that we ask annotators to compare two potentially distinct usages (in our case, dictionary defintions) and rate how similar they are. Because we use non-expert annotators, we switch to a 3-point scale; similar to (Erk et al., 2009), we also offer an option of not being able to decide.

3. Methodology

3.1. Pooling semi-experts

Developing a gold standard with expert annotators can be costly. An alternative for annotation for various NLP tasks is to collect non-expert annotations, i.e., crowdsourcing (Madnani et al., 2011; Wang et al., 2009; Snow et al., 2008), but the task of assigning sense mappings may be beyond the ability of most non-experts, even with training. We therefore pursue the strategy of collecting judgments from *semi-experts*, namely (computational) linguistics faculty and students at our universities (Muhonen and Purtonen, 2011). By surveying linguistics faculty and graduate students to align word senses, we thus target people who have at least a basic knowledge of semantics. We incur no costs, as participants are volunteers, and we thereby also mitigate the crowdsourcing problem of obtaining noisy answers (Laws et al., 2011), while at the same time being able to gather a number of annotators for a given sense.[2]

One limitation is the smaller potential pool of respondents than with crowdsourcing, more strictly limiting the amount of data that can be gathered. As an experiment into how resources align, however, the method is straightforward. One can use this methodology to develop a small data set, and/or to provide a platform for developing an experiment with crowdsourcing of non-experts or, in the other direction, to develop guidelines for expert annotators.

Our experiment sheds light on two questions: 1) How reliable are semi-experts at providing information pertinent to evaluating WSA systems? 2) How difficult is it to align WordNet and COBUILD? In both cases, we are interested in how much variability there is among respondents.

3.2. Word selection

We begin with basic words, relying on the Academic Word List (AWL). The AWL consists of 570 word families occurring most frequently over a range of academic texts, namely over 100 times in a 3.5 million word academic corpus.[3] These word families are indexed by a particular head word, e.g., *interpret* heads a list containing *interprets*, *interpreter*, *interpretation*, etc. Students who master the AWL thus greatly expand their vocabulary usage.

We select words with at least 3 WordNet (WN) senses, in order to obtain enough complexity to get a grasp on the general properties of alignment. We pick three types of words, representing a range of different COBUILD (CB) senses: 1) less senses than in WordNet; 2) (roughly) the same number of senses; and 3) more senses.[4] This gives us different

[2] More annotators can of course reduce any idiosyncracies arising from one person's data; see, for example, the discussion in (Erk and McCarthy, 2009).

[3] http://www.victoria.ac.nz

[4] We only have one instance of this (*indicate.v*); in general, COBUILD is less fine-grained than WordNet.

degrees of alignment *skewedness*, increasing the chances of seeing both zero/null mappings (i.e., where a sense in one inventory maps to nothing in the other) and multiple mappings. Despite being a small set, this break-down to some extent allows us to get a handle on word alignment across a diverse set of cases (cf. (Meyer and Gurevych, 2011)), just as (Erk and McCarthy, 2009) use eight lemmas to evaluate graded word sense assignment.[5] The nine selected words are in Table 1. In total, there are 63 WordNet and 35 COBUILD senses, incorporating both nouns and verbs.

	Word	WN	CB
Balanced	area.n	6	6
	indicate.v	5	6
	policy.n	3	3
Skewed	community.n	6	3
	involve.v	7	5
	job.n	12	4
	process.v	6	2
	require.v	4	2
	section.n	14	3

Table 1: Words selected for our experiment, including number of senses in each inventory

3.3. Survey design

Taking the 63 WordNet senses, the study consists of seven individual surveys with nine multiple-choice questions each. Each question is a WordNet sense, and the nine different words are distributed across the surveys. The question choices consist of all the COBUILD senses of a word (with examples), as in Figure 1. Each question includes examples of the sense; adding examples to the definition helps participants to more readily understand the sense. As shown, there are four options for each choice: *same meaning*, *related meaning*, *no relation*, and *unable to determine*. The last category is important, as it allows us to see how often participants had extreme difficulty in making a decision; such cases are the ones which would likely be the ones most in need of explicit guidelines.

We considered subdividing *related meaning* into specific cases, such as hyponymy, but kept it simple, to reduce cognitive load. Furthermore, we also considered not including *related meaning* at all, so as to better model a yes/no judgment task; however, this seemed not to match our own intuitions about the nature of the alignments, namely that they may be gradable (Erk and McCarthy, 2009; Erk et al., 2009) or may contain non-exact similarities (Meyer and Gurevych, 2011).

We use the WordNet sense as the question and the COBUILD senses as choices since we are ultimately interested in working in this direction, i.e., from a WordNet-based WSD system to COBUILD examples. However, alignments derived from the surveys can in principle work in either direction. In addition, presenting senses in a dictionary format (i.e. as definitions) is based on the purpose of the current study, in which we try to map sense definitions between dictionaries.

The final question of every survey is a question about participant confidence for all questions, using a Likert scale, as shown in Figure 2. In addition to the *unable to determine* cateogry, this helps us determine annotator ability and reliability for semi-experts.[6]

The surveys were administered to Linguistics faculty and students in the Departments of Linguistics and related fields at Georgetown University and Indiana University. Volunteers completed and anonymously submitted the surveys online. The surveys were administered via a free web service.[7] While this makes implementing such experiments feasible for researches in almost any context, there are distinct limitations, such as no being able to track the same user across different surveys.

4. Evaluation

Before delving into detailed evaluation, we can look at an example set of responses, as in Table 2. This is for the first WordNet sense (W1) of the noun *section*, which has 3 possible corresponding COBUILD senses. Sense 2 (C2) is a favorite, but C1 is also likely; C3 is divided, leaning towards not related.

This variability is typical of the responses, as we can see in Figure 3, where we sum the counts for each type of response for each word. We can also see the differing numbers of annotators in this graph, with *job.n*, for example, receiving more responses than *policy.n* in our experiment. For a word like *job.n*, the number of responses for *no relation* predominate, but for *community.n*, there are more *related meaning* instances. Most notably, as with the study in (Erk et al., 2009), respondents are clearly using not just the extreme categories (same/different), but are making great use of the *related meaning* category. Indeed, in total, *no relation* was the most popular answer (866 responses), followed closely by *related meaning* (828) and then *same meaning* (472); *Unable to determine* (146) was the least popular choice, but still accounted for 6.3% of the responses.

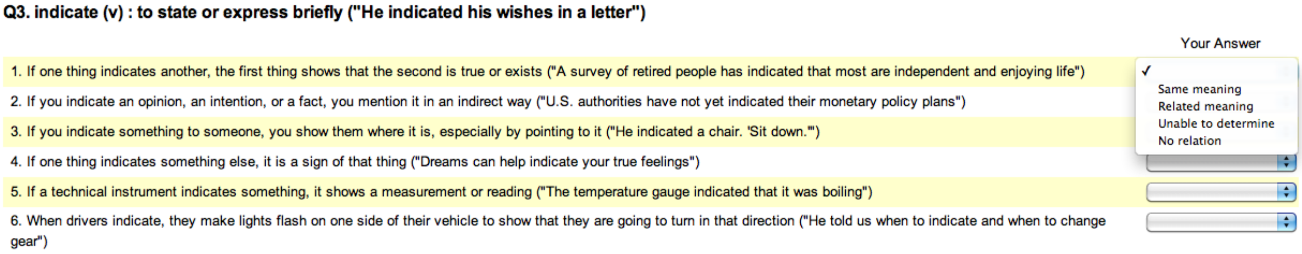
Turning to how well respondents agreed on their answers, when we calculate Fleiss' kappa to test interannotator agreement, we obtain a value of 0.18; according to (Landis and Koch, 1977), this is only "slight" agreement. This lack of agreement is not surprising if we look at participants' confidence in Table 3. Around 50% of WordNet senses result in confidence scores of 3 or below.

Part of the difficulty seems to lie in the fact that within each inventory, senses are related in complicated ways, sometimes causing confusion for annotators in mapping between them. For *community.n*, for example, the three COBUILD senses are:

[5] In the future, one can ensure selection across further criteria, including the so-called *Unique Beginner* of a word and location within the WordNet taxonomy (Niemann and Gurevych, 2011).

[6] Based on user feedback, for future surveys, we are placing a confidence rating directly after each question.

[7] `http://www.surveymonkey.com/`

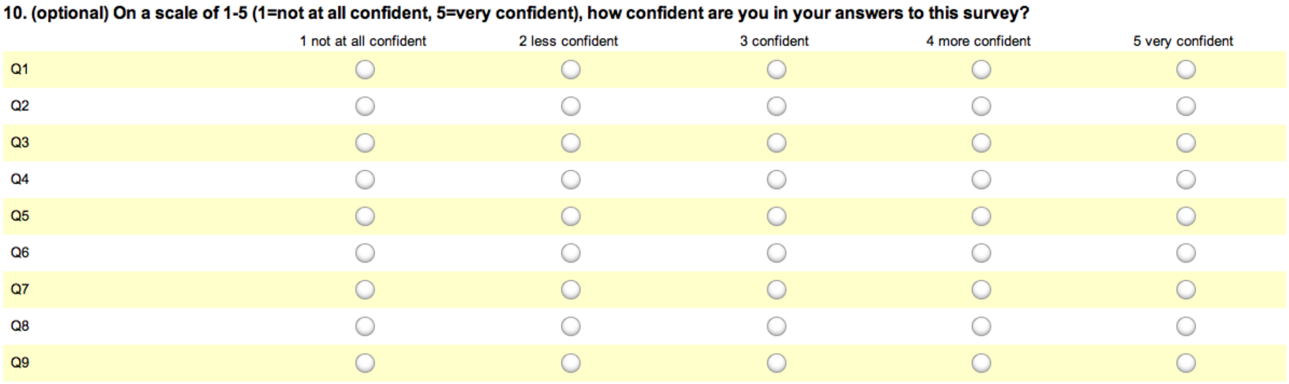


Q3. indicate (v) : to state or express briefly ("He indicated his wishes in a letter")

Your Answer

1. If one thing indicates another, the first thing shows that the second is true or exists ("A survey of retired people has indicated that most are independent and enjoying life")
2. If you indicate an opinion, an intention, or a fact, you mention it in an indirect way ("U.S. authorities have not yet indicated their monetary policy plans")
3. If you indicate something to someone, you show them where it is, especially by pointing to it ("He indicated a chair. 'Sit down.'")
4. If one thing indicates something else, it is a sign of that thing ("Dreams can help indicate your true feelings")
5. If a technical instrument indicates something, it shows a measurement or reading ("The temperature gauge indicated that it was boiling")
6. When drivers indicate, they make lights flash on one side of their vehicle to show that they are going to turn in that direction ("He told us when to indicate and when to change gear")

Figure 1: An example of a question and choices for one sense of *indicate.v*



10. (optional) On a scale of 1-5 (1=not at all confident, 5=very confident), how confident are you in your answers to this survey?

	1 not at all confident	2 less confident	3 confident	4 more confident	5 very confident
Q1	○	○	○	○	○
Q2	○	○	○	○	○
Q3	○	○	○	○	○
Q4	○	○	○	○	○
Q5	○	○	○	○	○
Q6	○	○	○	○	○
Q7	○	○	○	○	○
Q8	○	○	○	○	○
Q9	○	○	○	○	○

Figure 2: Confidence Scale

section (n) : a self-contained part of a larger composition ...	Same meaning	Related meaning	Unable to determine	No relation
1. A section of something is one of the parts into which it is divided ...	38.5% (5)	53.8% (7)	7.7% (1)	0.0% (0)
2. A section of an official document ... is one of the parts into which it is divided ...	76.9% (10)	15.4% (2)	0.0% (0)	7.7% (1)
3. A section is a diagram of something such as a building ...	0.0% (0)	38.5% (5)	7.7% (1)	53.8% (7)

Table 2: Response Analysis for one WordNet sense of *section.n*

1	2	3	4	5
0.21	0.97	1.98	1.63	1.70

Table 3: Average number of responses for each point on the confidence scale (1=not confident, 5 = very confident)

1. The community is all the people who live in a particular area or place.
2. A particular community is a group of people who are similar in some way.
3. Community is friendship between different people or groups, and a sense of having something in common.

When asked to align the WordNet sense of *common ownership*, then, this property can cut across all three defintions, as it seems to be describing a different way at looking at *community* completely.

Given the lack of agreement, an immediate question is: can we use these results to evaluate WSA systems? And if so, how? One answer is that the results should be used as weighted scores. That is, when evaluating measures such as precision and recall, instead of counting C2, for example, as a totally correct alignment for W1 of *section.n* (cf. Table 2), it counts as .769 of a correct alignment. One can see (Madnani et al., 2011) for such a proposal using binary crowdsourced data, and (Erk and McCarthy, 2009) for different measurements related to graded word senses.

An alternative is to seek whether we can obtain higher confidence in the way that the classes are used. To address this, we adjust our calculations by removing the *unable to determine* cases and combining *same* and *related* meanings. This reflects the fact that we may want to group them together for particular alignment uses; this gives a kappa of 0.24 (“fair” agreement). Again, the low agreement is not terribly surprising, given the low confidence reported ear-

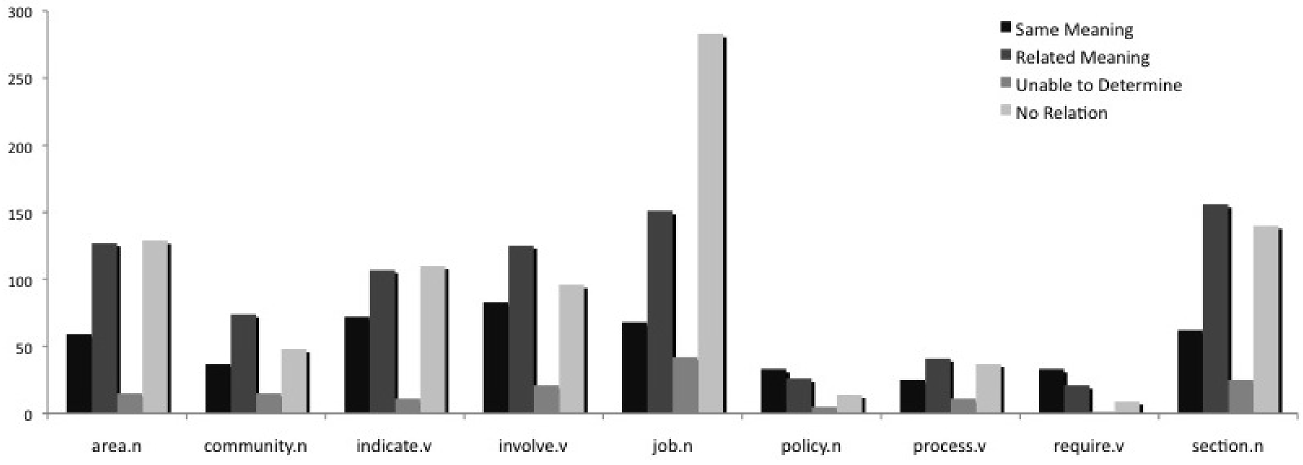


Figure 3: Number of times each answer was used for every word

lier, and it can indicate at least two things: 1) the task was not clear, or 2) these particular sense inventories are difficult to align. In the future, one may want to explore developing further guidelines, balancing this with the fact that volunteers are being used.

4.1. Converting responses to scores

We convert the responses into scores for evaluation, in order to quantify to what extent—according to the various annotators—the senses from the two inventories express the same meaning. Specifically, we assign a weight of 1 for *same meaning*, -1 for *no relation*, and 0 for *unable to determine*; thus, higher scores indicate a greater degree of "sameness."[8] For *related meaning*, we test different weights (α)—1, 0.5, and 0—reflecting differing degrees of their contribution towards a correct alignment. For example, for the W1-C1 mapping in table 2, we obtain: 12 $(= 5+7*1+0)$, 8.5 $(= 5+7*0.5+0)$, and 5 $(= 5+7*0+0)$, respectively.

Participants were not required to complete all surveys, so the number of responses per survey is different. thus, we normalize the scores by the number of respondents: in this case, with $\alpha = 0.5$, the score for W1-C1 is $\frac{8.5}{13} = 0.65$. For example, normalized scores for *involve.v* are in Table 4.

	C1	C2	C3	C4	C5
W1	0.083	0.5	0.333	0.083	0.25
W2	-0.063	0.75	0.688	0.5	0.125
W3	1	-0.143	-0.286	-0.143	-0.429
W4	0.611	-0.111	-0.444	-0.056	-0.389
W5	0.893	0.679	0.25	0.357	-0.071
W6	-0.455	0.227	0.227	0.409	-0.318
W7	-0.571	-0.143	0.286	-0.214	0

Table 4: Scores for *involve.v* ($\alpha = 0.5$)

4.2. Evaluating a basic WSA system

A simple word sense alignment system consists of running a basic WordNet-based WSD classifier on the COBUILD example sentences and averaging the scores, in order to find the best WordNet sense for a given COBUILD sense. Because we are interested in mapping from WordNet to a single-best COBUILD sense (see section 1.), we use these scores to take the most likely COBUILD sense for each WordNet sense (i.e., a 1-to-n mapping from WordNet to COBUILD). One could explore more robust alignment algorithms or take all senses above a given threshold to allow for many-to-many mappings, but this gives us a good starting point for testing evaluation data under the conditions we are interested in. We implement this by running SenseRelate::AllWords (SR::AW) (Pedersen and Kolhatkar, 2009) on the COBUILD sentences.

Using our scores, we perform two ways of counting different cases as correct alignments, namely counting: 1) all positive scores (unshaded cells of Table 4); or 2) only the top positive score for each WordNet sense (i.e., the highest score reading across a row). We could explore a graded notion of what counts as correct to calculate precision and recall (Erk and McCarthy, 2009), but we use our WSD system as a categorical one, returning *yes* or *no* for each alignment link. Thus, for present purposes, we convert our gold standard to categorical decisions; obviously, it can also be used for non-categorical evaluation. After defining a set of correct alignments, we calculate precision and recall of alignments in the usual way. The results are in Table 5, where we also report the number of senses which do not align to the other inventory.

For example, the system outputs (W1,C1), (W2,C2), (W3,C3), (W4,C5), (W5,C1), (W6,C2), (W7,C4) for *involve.v*. If all positive scores are correct alignments, the correct matches (for $\alpha = 0.5$) are (W1,C1), (W2,C2),

[8] One could also use normalized judgment scores as in (Erk and McCarthy, 2009). In our context, this means: *same*=2, *related*=1, *none*=0, and normalized score = $\frac{score}{2}$. Instead of ranging from -1 to 1, it ranges from 0 to 1, but shares the same basic intuition, especially for when $\alpha = 0$, putting *related meaning* exactly halfway between the others.

		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 0$	
		AP	TP	AP	TP	AP	TP
area.n	P	0.67	0.33	0.33	0.33	0.17	0.17
	R	0.21	0.33	0.18	0.33	0.14	0.20
community.n	P	0.83	0.50	0.67	0.50	0.67	0.50
	R	0.38	0.50	0.67	0.50	0.57	0.60
indicate.v	P	0.60	0.20	0.20	0.20	0.20	0.20
	R	0.16	0.25	0.09	0.25	0.14	0.25
involve.v	P	0.71	0.14	0.57	0.29	0.29	0.14
	R	0.20	0.14	0.21	0.29	0.18	0.17
job.n	P	0.42	0.25	0.33	0.17	0.17	0.08
	R	0.21	0.30	0.22	0.22	0.17	0.14
policy.n	P	1.00	0.67	1.00	0.67	1.00	0.67
	R	0.38	0.67	0.43	0.67	0.60	0.67
process.n	P	0.50	0.50	0.50	0.50	0.50	0.50
	R	0.43	0.50	0.43	0.50	0.60	0.60
require.v	P	1.00	0.75	1.00	0.75	0.75	0.75
	R	0.50	0.75	0.57	0.75	0.60	0.75
section.n	P	0.86	0.50	0.64	0.36	0.43	0.21
	R	0.57	0.50	0.56	0.42	0.55	0.38
un-aligned	WN	**4**	5	8	8	19	19
senses	CB	**0**	6	**0**	5	2	6

Table 5: Precision & recall of words using SR::AW (AP=all positives, TP=top positive), plus number of un-aligned senses

(W5,C1) and (W6,C2). Consequently, precision is $\frac{4}{7} = 0.57$ and recall $\frac{4}{19} = 0.21$.

With $\alpha = 1$, related senses are counted fully correct, meaning the system will match more, giving higher precision. Likewise, $\alpha = 0$ gives fewer alignments, producing generally higher recall and more unalignment between the inventories.

These results do not indicate the best evaluation; they simply illustrate how the treatment of related senses affects the results. The all positive (AP), $\alpha = 1$ evaluation, for instance, indicates how far off a system is from any correct answer, while, on the other side of the spectrum, the top positive (TP), $\alpha = 0$ evaluation indicates how well the best senses are being found. To display sense-specific examples for learners, we will want evaluations across the spectrum to know how often learners will be presented with related examples, as opposed to exact matches.

5. Summary and Outlook

We have examined constructing a database of alignments of word senses between two sense inventories, specifically WordNet and COBUILD, by pooling the judgments of semi-experts. Using online surveys, we presented a sense of a target word from one dictionary with senses from the other dictionary, asking for judgments of relatedness. Specifically, we have shown: 1) It is difficult for semi-experts to agree upon correct alignments, showing that the task is difficult, and it would seem to be infeasible for, e.g., crowdsourcing of non-experts. 2) Despite this, such data can be used to gauge accuracy of WSA systems, depending upon how much *related meaning* one wishes to capture in the alignments.

Currently, we are obtaining more data from more surveys, focusing on words which are relevant to the system for reading assistance we are building. At the same time, we are investigating different ways to use the annotator judgments to evaluate WSA systems, building from work such as (Erk and McCarthy, 2009).

Acknowledgments

We wish to thank the survey respondents for their help, as well as three anonymous reviewers for their feedback.

6. References

Sue Atkins. 1993. Tools for computer-aided corpus lexicography: the hector proje ct. *Acta Linguistica Hungarica*, 41:5–72.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore, August.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore, August.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Samuel Fernando and Mark Stevenson. 2010. Aligning wordnet synsets and wikipedia articles. In *Proceedings of the AAAI Workshop on Collaboratively-Built Knowledge Sources and Artificial Intelligence*, pages 48–50, Athens, GA.

M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2006. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the 9th International Conference on Spoken Language Processing*.

Nancy Ide and Jean Véronis. 1990. Mapping dictionaries: A spreading activation approach. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary*, pages 52–64, Waterloo.

Kevin Knight and Steve K. Luk. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of AAAI-94*, Seattle.

Oi Yee Kwong. 1998. Aligning WordNet with additional lexical resources. In *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 73–79, Montreal.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Florian Laws, Christian Scheible, and Hinrich Schutze. 2011. Active learning with amazon mechanical turk. In *Proceedings of EMNLP-11*, Edinburgh.

Nitin Madnani, Martin Chodorow, Joel Tetreault, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 508–513, Portland, OR, June.

Christian M. Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892, Chiang Mai, Thailand, November.

Kristiina Muhonen and Tanja Purtonen. 2011. Creating a dependency syntactic treebank: Towards intuitive language modeling. In *Proceedings of DepLing-11*, Barcelona.

Vivi Nastase and Stan Szpakowicz. 2001. Word-sense disambiguation in Roget's Thesaurus using WordNet. In *Proceedings of the Workshop on WordNet and other lexical resources*, Pittsburgh.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of ACL-10*, pages 216–225, Uppsala, Sweden, July.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the* 4^{th} *International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic.

Roberto Navigli. 2006. Reducing the granularity of a computational lexicon via an automatic mapping to a coarse-grained sense inventory. In *Proceedings of LREC 2006*, Genova.

Roberto Navigli. 2009. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

John Nerbonne and Petra Smit. 1996. GLOSSER-RuG: in support of reading. In *Proceedings of COLING-96*.

Elisabeth Niemann and Iryna Gurevych. 2011. The peoples web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Oxford.

Martha Palmer, Hoa Trang Dang, and Joseph Rosenzweig. 2000. Sense tagging the Penn Treebank. In *Proceedings of LREC-00*, Athens.

Ted Pedersen and Varada Kolhatkar. 2009. WordNet::SenseRelate::AllWords - a broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of the North American Chapter of the Association for Computational Linguistics- Human Language Technology 2009 Conference*, Boulder, CO.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *Advances in Web Intelligence*, pages 380–386.

John Sinclair, editor. 2006. *Collins COBUILD Advanced Lerner's English Dictionary*. Harper Collins.

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. *Proceedings of EMNLP-08*.

Antonio Toral, Óscar Ferrández, Eneko Agirre, and Rafael Mu noz. 2009. A study on linking and disambiguating wikipedia categories to wordnet using text similarity. In *Proceedings of RANLP 09*, Borovets, Bulgaria.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2009. Perspectives on crowdshourcing annotations for natural language processing. *Language Resources and Evaluation*.

Elisabeth Wolf and Iryna Gurevych. 2010. Aligning sense inventories in wikipedia and wordnet. In *Proceedings of the First Workshop on Automated Knowledge Base Construction*, pages 24–28, Grenoble, France.