Combining Orthogonal Monolingual and Multilingual Sources of Evidence for All Words WSD

Weiwei Guo
Computer Science Department
Columbia University
New York, NY, 10115
weiwei@cs.columbia.edu

Mona Diab
Center for Computational Learning Systems
Columbia University
New York, NY, 10115
mdiab@ccls.columbia.edu

Abstract

Word Sense Disambiguation remains one of the most complex problems facing computational linguists to date. In this paper we present a system that combines evidence from a monolingual WSD system together with that from a multilingual WSD system to yield state of the art performance on standard All-Words data sets. The monolingual system is based on a modification of the graph based state of the art algorithm In-Degree. The multilingual system is an improvement over an All-Words unsupervised approach, SALAAM. SALAAM exploits multilingual evidence as a means of disambiguation. In this paper, we present modifications to both of the original approaches and then their combination. We finally report the highest results obtained to date on the SENSEVAL 2 standard data set using an unsupervised method, we achieve an overall F measure of 64.58 using a voting scheme.

1 Introduction

Despite advances in natural language processing (NLP), Word Sense Disambiguation (WSD) is still considered one of the most challenging problems in the field. Ever since the field's inception, WSD has been perceived as one of the central problems in NLP. WSD is viewed as an enabling technology that could potentially have far reaching impact on NLP applications in general. We are starting to see the beginnings of a positive effect of WSD in NLP applications such as Machine Translation (Carpuat and Wu, 2007; Chan et al., 2007).

Advances in WSD research in the current millennium can be attributed to several key factors: the availability of large scale computational lexical resources such as WordNets (Fellbaum, 1998; Miller, 1990), the availability of large scale corpora, the existence and dissemination of standardized data sets over the past 10 years through different testbeds such as SENSEVAL and SEMEVAL competitions,[1] devising more robust computing algorithms to handle large scale data sets, and simply advancement in hardware machinery.

In this paper, we address the problem of WSD of all content words in a sentence, All-Words data. In this framework, the task is to associate all tokens with their contextually relevant meaning definitions from some computational lexical resource. Our work hinges upon combining two high quality WSD systems that rely on essentially different sources of evidence. The two WSD systems are a monolingual system `RelCont` and a multilingual system `TransCont`. `RelCont` is an enhancement on an existing graph based algorithm, In-Degree, first described in (Navigli and Lapata, 2007). `TransCont` is an enhancement over an existing approach that leverages multilingual evidence through projection, `SALAAM`, described in detail in (Diab and Resnik, 2002). Similar to the leveraged systems, the current combined approach is unsupervised, namely it does not rely on training data from the onset. We show that by combining both sources of evidence, our approach yields the highest performance for an unsupervised system to date on standard All-Words data sets.

This paper is organized as follows: Section 2 delves into the problem of WSD in more detail; Section 3 explores some of the relevant related work; in Section 4, we describe the two WSD systems in some detail emphasizing the improvements to the basic systems in addition to a description of our combination approach; we present our experimental set up and results in Section 5; we discuss the results and our overall observations with error analysis in Section 6; Finally, we con-

[1] http://www.semeval.org

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1542–1551,
Uppsala, Sweden, 11-16 July 2010.

clude in Section 7.

2 Word Sense Disambiguation

The definition of WSD has taken on several different practical meanings in recent years. In the latest SEMEVAL 2010 workshop, there are 18 tasks defined, several of which are on different languages, however we recognize the widening of the definition of the task of WSD. In addition to the traditional All-Words and Lexical Sample tasks, we note new tasks on word sense discrimination (no sense inventory needed, the different senses are merely distinguished), lexical substitution using synonyms of words as substitutes both monolingually and multilingually, as well as meaning definitions obtained from different languages namely using words in translation.

Our paper is about the classical All-Words (AW) task of WSD. In this task, all content bearing words in running text are disambiguated from a static lexical resource. For example a sentence such as 'I walked by the bank and saw many beautiful plants there.' will have the verbs 'walked, saw', the nouns 'bank, plants', the adjectives 'many, beautiful', and the adverb 'there', be disambiguated from a standard lexical resource. Hence, using WordNet,[2] 'walked' will be assigned the corresponding meaning definitions of: *t*o use one's feet to advance; to advance by steps, 'saw' will be assigned the meaning definition of: *t*o perceive by sight or have the power to perceive by sight, the noun 'bank' will be assigned the meaning definition of: *s*loping land especially the slope beside a body of water, and so on.

3 Related Works

Many systems over the years have been proposed for the task. A thorough review of the state of the art through the late 1990s (Ide and Veronis, 1998) and more recently in (Navigli, 2009). Several techniques have been used to tackle the problem ranging from rule based/knowledge based approaches to unsupervised and supervised machine learning techniques. To date, the best approaches that solve the AW WSD task are supervised as illustrated in the different SenseEval and SEMEVAL AW task (Palmer et al., 2001; Snyder and Palmer, 2004; Pradhan et al., 2007).

In this paper, we present an unsupervised combination approach to the AW WSD problem that relies on WN similarity measures in conjunction with evidence obtained through exploiting multilingual evidence. We will review the closely relevant related work on which this current investigation is based.[3]

4 Our Approach

Our current investigation exploits two basic unsupervised approaches that perform at state-of-the-art for the AW WSD task in an unsupervised setting. Crucially the two systems rely on different sources of evidence allowing them to complement each other to a large extent leading to better performance than for each system independently. Given a target content word and co-occurring contextual clues, the monolingual system `RelCont` attempts to assign the approporiate meaning definition to the target word. Such words by definition are semantically related words. `TransCont`, on the other hand, is the multilingual system. `TransCont` defines the notion of context in the translational space using a foreign word as a filter for defining the contextual content words for a given target word. In this multilingual setting, all the words that are mapped to (aligned with) the same orthographic form in a foreign language constitute the context. In the next subsections we describe the two approaches `RelCont` and `TransCont` in some detail, then we proceed to describe two combination methods for the two approaches: `MERGE` and `VOTE`.

4.1 Monolingual System `RelCont`

`RelCont` is based on an extension of a state-of-the-art WSD approach by (Sinha and Mihalcea, 2007), henceforth (SM07). In the basic SM07 work, the authors combine different semantic similarity measures with different graph based algorithms as an extension to work in (Mihalcea, 2005). Given a sequence of words $W = \{w_1, w_2 ... w_n\}$, each word w_i with several senses $\{s_{i1}, s_{i2} ... s_{im}\}$. A graph G = (V,E) is defined such that there exists a vertex v for each sense. Two senses of two different words may be connected by an edge e, depending on their distance. That two senses are connected suggests they should have influence on each other, accordingly a maximum

[2] http://wordnet.princeton.edu

[3] We acknowledge the existence of many research papers that tackled the AW WSD problem using unsupervised approaches, yet for lack of space we will not be able to review most of them.

allowable distance is set. They explore 4 different graph based algorithms. The highest yielding algorithm in their work is the In-Degree algorithm combining different WN similarity measures depending on POS. They used the Jiang and Conrath (JCN) (Jiang and Conrath., 1997) similarity measure within nouns, the Leacock & Chodorow (LCH) (Leacock and Chodorow, 1998) similarity measure within verbs, and the Lesk (Lesk, 1986) similarity measure within adjectives, within adverbs, and among different POS tag pairings. They evaluate their work against the SENSEVAL 2 AW test data (SV2AW). They tune the parameters of their algorithm – namely, the normalization ratio for some of these measures – on the SENSEVAL 3 data set. They report a state-of-the-art unsupervised system that yields an overall performance across all AW POS sets of 57.2%.

In our current work, we extend the SM07 work in some interesting ways. A detailed narrative of our approach is described in (Guo and Diab, 2009). Briefly, we focus on the In-Degree graph based algorithm since it is the best performer in the SM07 work. The In-Degree algorithm presents the problem as a weighted graph with senses as nodes and the similarity between senses as weights on edges. The In-Degree of a vertex refers to the number of edges incident on that vertex. In the weighted graph, the In-Degree for each vertex is calculated by summing the weights on the edges that are incident on it. After all the In-Degree values for each sense are computed, the sense with maximum value is chosen as the final sense for that word.

In this paper, we use the In-Degree algorithm while applying some modifications to the basic similarity measures exploited and the WN lexical resource tapped into. Similar to the original In-Degree algorithm, we produce a probabilistic ranked list of senses. Our modifications are described as follows:

JCN for Verb-Verb Similarity In our implementation of the In-Degree algorithm, we use the JCN similarity measure for both Noun-Noun similarity calculation similar to SM07. However, different from SM07, instead of using LCH for Verb-Verb similarity, we use the JCN metric as it yields better performance in our experimentations.

Expand Lesk Following the intuition in (Pedersen et al., 2005), henceforth (PEA05), we expand the basic Lesk similarity measure to take into account the glosses for all the relations for the synsets on the contextual words and compare them with the glosses of the target word senses, therefore going beyond the is-a relation. We exploit the observation that WN senses are too fine-grained, accordingly the neighbors would be slightly varied while sharing significant semantic meaning content. To find similar senses, we use the relations: hypernym, hyponym, similar attributes, similar verb group, pertinym, holonym, and meronyms.[4] The algorithm assumes that the words in the input are POS tagged. In PEA05, the authors retrieve all the relevant neighbors to form a bag of words for both the target sense and the surrounding senses of the context words, they specifically focus on the Lesk similarity measure. In our current work, we employ the neighbors in a disambiguation strategy using different similarity measures one pair at a time. Our algorithm takes as input a target sense and a sense pertaining to a word in the surrounding context, and returns a sense similarity score. We do not apply the WN relations expansion to the target sense. It is only applied to the contextual word.[5]

For the monolingual system, we employ the same normalization values used in SM07 for the different similarity measures. Namely for the Lesk and Expand-Lesk, we use the same cut-off value of 240, accordingly, if the Lesk or Expand-Lesk similarity value returns $0 <= 240$ it is converted to a real number in the interval [0,1], any similarity over 240 is by default mapped to 1. We will refer to the Expand-Lesk with this threshold as Lesk2. We also experimented with different thresholds for the Lesk and Expand-Lesk similarity measure using the SENSEVAL 3 data as a tuning set. We found that a cut-off threshold of 40 was also useful. We will refer to this variant of Expand-Lesk with a cut off threshold of 40 as Lesk3. For JCN, similar to SM07, the values are from 0.04 to 0.2, we mapped them to the interval [0,1]. We did not run any calibration studies beyond the what was reported in SM07.

[4] In our experiments, we varied the number of relations to employ and they all yielded relatively similar results. Hence in this paper, we report results using all the relations listed above.

[5] We experimented with expanding both the contextual sense and the target sense and we found that the unreliability of some of the relations is detrimental to the algorithm's performance. Hence we decided empirically to expand only the contextual word.

SemCor Expansion of WN A part of the `RelCont` approach relies on using the Lesk algorithm. Accordingly, the availability of glosses associated with the WN entries is extremely beneficial. Therefore, we expand the number of glosses available in WN by using the SemCor data set, thereby adding more examples to compare. The SemCor corpus is a corpus that is manually sense tagged (Miller, 1990).[6] In this expansion, depending on the version of WN, we use the sense-index file in the WN Database to convert the SemCor data to the appropriate version sense annotations. We augment the sense entries for the different POS WN databases with example usages from SemCor. The augmentation is done as a look up table external to WN proper since we did not want to dabble with the WN offsets. We set a cap of 30 additional examples per synset. We used the first 30 examples with no filtering criteria. Many of the synsets had no additional examples. WN1.7.1 comprises a total of 26875 synsets, of which 25940 synsets are augmented with SemCor examples.[7]

4.2 Multilingual System `TransCont`

`TransCont` is based on the WSD system SALAAM (Diab and Resnik, 2002), henceforth (DR02). The SALAAM system leverages word alignments from parallel corpora to perform WSD. The SALAAM algorithm exploits the word correspondence cross linguistically to tag word senses on words in running text. It relies on several underlying assumptions. The first assumption is that senses of polysemous words in one language could be lexicalized differently in other languages. For example, 'bank' in English would be translated as *banque* or *rive de fleuve* in French, depending on context. The other assumption is that if Language 1 (L1) words are translated to the same orthographic form in Language 2 (L2), then they share the some element of meaning, they are semantically similar.[8]

The SALAAM algorithm can be described as follows. Given a parallel corpus of L1-L2 that is sentence and word aligned, group all the word types in L1 that map to same word in L2 creating clusters referred to as *t*ypesets. Then perform disambiguation on the typeset clusters using WN. Once senses are identified for each word in the cluster, the senses are propagated back to the original word instances in the corpus. In the SALAAM algorithm, the disambiguation step is carried out as follows: within each of these target sets consider all possible sense tags for each word and choose sense tags informed by semantic similarity with all the other words in the whole group. The algorithm is a greedy algorithm that aims at maximizing the similarity of the chosen sense across all the words in the set. The SALAAM disambiguation algorithm used the noun groupings (NounGroupings) algorithm described in DR02. The algorithm applies disambiguation within POS tag. The authors report only results on the nouns only since NounGroupings heavily exploits the hierarchy structure of the WN noun taxonomy, which does not exist for adjectives and adverbs, and is very shallow for verbs.

Essentially SALAAM relies on variability in translation as it is important to have multiple words in a typeset to allow for disambiguation. In the original SALAAM system, the authors automatically translated several balanced corpora in order to render more variable data for the approach to show it's impact. The corpora that were translated are: the WSJ, the Brown corpus and all the SENSEVAL data. The data were translated to different languages (Arabic, French and Spanish) using state of art MT systems. They employed the automatic alignment system GIZA++ (Och and Ney, 2003) to obtain word alignments in a single direction from L1 to L2.

For `TransCont` we use the basic SALAAM approach with some crucial modifications that lead to better performance. We still rely on parallel corpora, we extract typesets based on the intersection of word alignments in both alignment directions using more advanced GIZA++ machinery. In contrast to DR02, we experiment with all four POS: Verbs (V), Nouns (N), Adjectives (A) and Adverbs (R). Moreover, we modified the underlying disambiguation method on the typesets. We still employ WN similarity, however, we do not use the NounGroupings algorithm. Our disambiguation method relies on calculating the sense pair similarity exhaustively across all the

[6] Using SemCor in this setting to augment WN does hint of using supervised data in the WSD process, however, since our approach does not rely on training data and SemCor is not used in our algorithm directly to tag data, but to augment a rich knowledge resource, we contend that this does not affect our system's designation as an unsupervised system.

[7] Some example sentences are repeated across different synsets and POS since the SemCor data is annotated as an All-Words tagged data set.

[8] We implicitly make the underlying simplifying assumption that the L2 words are less ambiguous than the L1 words.

word types in a typeset and choosing the combination that yields the highest similarity. We experimented with all the WN similarity measures in the WN similarity package.[9] We also experiment with Lesk2 and Lesk3 as well as other measures, however we do not use SemCor examples with `TransCont`. We found that the best results are yielded using the Lesk2/Lesk3 similarity measure for N, A and R POS tagsets, while the Lin and JCN measures yield the best performance for the verbs. In contrast to the DR02 approach, we modify the internal WSD process to use the `In-Degree` algorithm on the typeset, so each sense obtains a confidence, and the sense(s) with the highest confidences are returned.

4.3 Combining `RelCont` and `TransCont`

Our objective is to combine the different sources of evidence for the purposes of producing an effective overall global WSD system that is able to disambiguate all content words in running text. We combine the two systems in two different ways.

4.3.1 MERGE

In this combination scheme, the words in the typeset that result from the `TransCont` approach are added to the context of the target word in the `RelCont` approach. However the typeset words are not treated the same as the words that come from the surrounding context in the `In-Degree` algorithm as we recognize that words that are yielded in the typesets are semantically similar in terms of content rather than being co-occurring words as is the case for contextual words in RelCont. Heeding this difference, we proceed to calculate similarity for words in the typesets using different similarity measures. In the case of noun-noun similarity, in the original `RelCont` experiments we use JCN, however with the words present in the `TransCont` typesets we use one of the Lesk variants, Lesk2 or Lesk3. Our observation is that the JCN measure is relatively coarser grained, compared to Lesk measures, therefore it is sufficient in case of lexical relatedness therefore works well in case of the context words. Yet for the words yielded in the `TransCont` typesets a method that exploits the underlying rich relations in the noun hierarchy captures the semantic similarity more aptly. In the case of verbs we still maintain the JCN similarity as it most effective given the shallowness of the verb hierarchy and the inherent nature of the verbal synsets which are differentiated along syntactic rather than semantic dimensions. We employ the Lesk algorithm still with A-A and R-R similarity and when comparing across different POS tag pairings.

4.3.2 VOTE

In this combination scheme, the output of the global disambiguation system is simply an intersection of the two outputs from the two underlying systems `RelCont` and `TransCont`. Specifically, we sum up the confidence ranging from 0 to 1 of the two system `In-Degree` algorithm outputs to obtain a final confidence for each sense, choosing the sense(s) that yields the highest confidences. The fact that `TransCont` uses `In-Degree` internally allows for a seamless integration.

5 Experiments and Results

5.1 Data

The parallel data we experiment with are the same standard data sets as in (Diab and Resnik, 2002), namely, Senseval 2 English AW data sets (SV2AW) (Palmer et al., 2001), and Seneval 3 English AW (SV3AW) data set. We use the true POS tag sets in the test data as rendered in the Penn Tree Bank.[10] We present our results on WordNet 1.7.1 for ease of comparison with previous results.

5.2 Evaluation Metrics

We use the `scorer2` software to report fine-grained (P)recision and (R)ecall and (F)-measure.

5.3 Baselines

We consider here several baselines. 1. A random baseline (RAND) is the most appropriate baseline for an unsupervised approach.2. We include the most frequent sense baseline (MFBL), though we note that we consider the most frequent sense or first sense baseline to be a supervised baseline since it depends crucially on SemCor in ranking the senses within WN.[11] 3. The SM07 results as a

[9] http://wn-similarity.sourceforge.net/

[10] We exclude the data points that have a tag of "U" in the gold standard for both baselines and our system.

[11] From an application standpoint, we do not find the first sense baseline to be of interest since it introduces a strong level of uniformity – removing semantic variability – which is not desirable. Even if the first sense achieves higher results in data sets, it is an artifact of the size of the data and the very limited number of documents under investigation.

monolingual baseline. 4. The DR02 results as the multilingual baseline.

5.4 Experimental Results

5.4.1 RelCont

We present the results for 4 different experimental conditions for RelCont: JCN-V which uses JCN instead of LCH for verb-verb similarity comparison, we consider this our base condition; +ExpandL is adding the Lesk Expansion to the base condition, namely Lesk2;[12] +SemCor adds the SemCor expansion to the base condition; and finally +ExpandL_SemCor, adds the latter both conditions simultaneously. Table 1 illustrates the obtained results for the SV2AW using WordNet 1.7.1 since it is the most studied data set and for ease of comparison with previous studies. We break the results down by POS tag (N)oun, (V)erb, (A)djective, and Adve(R)b. The coverage for SV2AW is 98.17% losing some of the verb and adverb target words.

Our overall results on all the data sets clearly outperform the baseline as well as state-of-the-art performance using an unsupervised system (SM07) in overall f-measure across all the data sets. We are unable to beat the most frequent baseline (MFBL) which is obtained using the first sense. However MFBL is a supervised baseline and our approach is unsupervised. Our implementation of SM07 is slightly higher than those reported in (Sinha and Mihalcea, 2007) (57.12%) is probably due to the fact that we do not consider the items tagged as "U" and also we resolve some of the POS tag mismatches between the gold set and the test data. We note that for the SV2AW data set our coverage is not 100% due to some POS tag mismatches that could not have been resolved automatically. These POS tag problems have to do mainly with multiword expressions. In observing the performance of the overall RelCont, we note that using JCN for verbs clearly outperforms using the LCH similarity measure. Using SemCor to augment WN examples seems to have the biggest impact. Combining SemCor with ExpandL yields the best results.

Observing the results yielded per POS in Table 1, ExpandL seems to have the biggest impact on the Nouns only. This is understandable since the noun hierarchy has the most dense relations and the most consistent ones. SemCor augmentation of WN seemed to benefit all POS significantly except for nouns. In fact the performance on the nouns deteriorated from the base condition JCN-V from 68.7 to 68.3%. This maybe due to inconsistencies in the annotations of nouns in SemCor or the very fine granularity of the nouns in WN. We know that 72% of the nouns, 74% of the verbs, 68.9% of the adjectives, and 81.9% of the adverbs directly exploited the use of SemCor augmented examples. Combining SemCor and ExpandL seems to have a positive impact on the verbs and adverbs, but not on the nouns and adjectives. These trends are not held consistently across data sets. For example, we see that SemCor augmentation helps all POS tag sets over using ExpandL alone or even when combined with SemCor. We note the similar trends in performance for the SV3AW data.

Compared to state of the art systems, RelCont with an overall F-measure performance of 62.13% outperforms the best unsupervised system of 57.5% UNED-AW-U2 for SV2 (Navigli, 2009). It is worth noting that it is higher than several of the supervised systems. Moreover, RelCont yields better overall results on SV3 at 59.87 compared to the best unsupervised system IRST-DDD-U which yielded an F-measure of 58.3% (Navigli, 2009).

5.4.2 TransCont

For the TransCont results we illustrate the original SALAAM results as our baseline. Similar to the DR02 work, we actually use the same SALAAM parallel corpora comprising more than 5.5M English tokens translated using a single machine translation system GlobalLink. Therefore our parallel corpus is the French English translation condition mentioned in DR02 work as FrGl. We have 4 experimental conditions: FRGL using Lesk2 for all POS tags in the typeset disambiguation (Lesk2); FRGL using Lesk3 for all POS tags (Lesk3); using Lesk3 for N, A and R but LIN similarity measure for verbs (Lesk3_Lin); using Lesk3 for N, A and R but JCN for verbs (Lesk3_JCN).

In Table 3 we note the the Lesk3_JCN followed immediately by Lesk3_Lin yield the best performance. The trend holds for both SV2AW and SV3AW. Essentially our new implementation of the multilingual system significantly outperforms the original DR02 implementation for all experimental conditions.

[12] Using Lesk3 yields almost the same results

Condition	N	V	A	R	Global F Measure
RAND	43.7	21	41.2	57.4	39.9
MFBL	71.8	41.45	67.7	81.8	65.35
SM07	68.7	33.01	65.2	63.1	59.2
JCN-V	68.7	35.46	65.2	63.1	59.72
+ExpandL	**70.2**	35.86	65.4	62.45	60.48
+SemCor	68.5	**38.66**	**69.2**	67.75	61.79
+ExpandL_SemCor	69.0	**38.66**	68.8	**69.45**	**62.13**

Table 1: `RelCont` F-measure results per POS tag per condition for SV2AW using WN 1.7.1.

Condition	N	V	A	R	Global F Measure
RAND	39.67	19.34	41.85	92.31	32.97
MFBL	70.4	54.15	66.7	92.88	63.96
SM07	60.9	43.4	57	92.88	53.98
JCN-V	60.9	48.5	57	92.88	55.87
+ExpandL	59.9	48.55	57.95	92.88	55.62
+SemCor	**66**	48.95	**65.55**	92.88	**59.87**
+ExpandL_SemCor	65	**49.2**	**65.55**	92.88	59.52

Table 2: `RelCont` F-measure results per POS tag per condition for SV3AW using WN 1.7.1.

5.4.3 Global Combined WSD

In this section we present the results of the global combined WSD system. All the combined experimental conditions have the same percentage coverage.[13] We present the results combining using MERGE and using VOTE. We have chosen 4 baseline systems: (1) SM07; (2) the our baseline monolingual system using JCN for verb-verb comparisons (RelCont-BL), so as to distinguish the level of improvement that could be attributed to the multilingual system in the combination results; as well as (3) and (4) our best individual system results from `RelCont` (ExpandL_SemCor) referred to in the tables below as (RelCont-Final) and `TransCont` using the best experimental condition (Lesk3_JCN). Table 5 and 6 illustrates the overall performance of our combined approach.

In Table 5 we note that the combined conditions outperform the two base systems independently, using `TransCont` is always helpful for any of the 3 monolingual systems, no matter we use VOTE or MERGE. In general the trend is that VOTE outperforms MERGE, however they exhibit different behaviors with respect to what works for each POS.

In Table 6 the combined result is not always better than the corresponding monolingual system. When applying to our baseline monolingual system, the combined result is still better. However, we observed worse results for ExpandL_Semcor, RelCont-Final. There may be 2 main reasons for the loss: (1) SV3 is the tuning set in SM07, and we inherit the thresholds for similarity metrics from that study. Accordingly, an overfitting of the thresholds is probably happening in this case; (2) `TransCont` results are not good enough on the SV3AW data. Comparing the `RelCont` and `TransCont` system results, we find a drop in f-measure of -1.37% in SV2AW, in contrast to a much larger drop in performance for the SV3AW data set where the drop in performance is -6.38% when comparing `RelCont-BL` to `TransCont` and nearly -10% comparing against `RelCont-Final`.

6 Discussion

We looked closely at the data in the combined conditions attempting to get a feel for the data and understand what was captured and what was not. Some of the good examples that are captured in the combined system that are not tagged in `RelCont` is the case of *ringer* in *Like most of the other 6,000 churches in Britain with sets of bells , St. Michael once had its own " band " of ringers , who would herald every Sunday morning and evening service* .. The `RelCont` answer is *ringer* sense number 4: *(horseshoes) the successful throw of a horseshoe*

[13] We do not back off in any of our systems to a default sense, hence the coverage is not at a 100%.

Condition	N	V	A	R	Global F Measure
RAND	43.7	21	41.2	57.4	39.9
DR02-FRGL	54.5				
SALAAM	65.48	31.77	56.87	67.4	57.23
Lesk2	67.05	30	59.69	68.01	57.27
Lesk3	67.15	30	60.2	68.01	57.41
Lesk3_Lin	67.15	29.27	**60.2**	**68.01**	57.61
Lesk3_JCN	**67.15**	**33.88**	**60.2**	**68.01**	**58.35**

Table 3: `TransCont` F-measure results per POS tag per condition for SV2AW using WN 1.7.1.

Condition	N	V	A	R	Global F Measure
RAND	39.67	19.34	41.85	92.31	32.93
SALAAM	52.42	29.27	54.14	88.89	45.63
Lesk2	53.57	33.58	53.63	88.89	47
Lesk3	53.77	33.30	56.48	88.89	47.5
Lesk3_Lin	53.77	29.24	56.48	88.89	46.37
Lesk3_JCN	53.77	38.43	56.48	88.89	**49.29**

Table 4: `TransCont` F-measure results per POS tag per condition for SV3AW using WN 1.7.1.

or quoit so as to encircle a stake or peg. When the merged system is employed we see the correct sense being chosen as sense number 1 in the MERGE condition: defined in WN as *a person who rings church bells (as for summoning the congregation)* resulting from a corresponding translation into French as *s*onneur.

We did some basic data analysis on the items we are incapable of capturing. Several of them are cases of metonymy in examples such as "the English are known...", the sense of *English* here is clearly in reference to the people of England, however, our WSD system preferred the language sense of the word. These cases are not gotten by any of our systems. If it had access to syntactic/semantic roles we assume it could capture that this sense of the word entails volition for example. Other types of errors resulted from the lack of a way to explicitly identify multiwords.

Looking at the performance of `TransCont` we note that much of the loss is a result of the lack of variability in the translations which is a key factor in the performance of the algorithm. For example for the 157 adjective target test words in SV2AW, there was a single word alignment for 51 of the cases, losing any tagging for these words.

7 Conclusions and Future Directions

In this paper we present a framework that combines orthogonal sources of evidence to create a state-of-the-art system for the task of WSD disambiguation for AW. Our approach yields an overall global F measure of 64.58 for the standard SV2AW data set combining monolingual and multilingual evidence. The approach can be further refined by adding other types of orthogonal features such as syntactic features and semantic role label features. Adding SemCor examples to `TransCont` should have a positive impact on performance. Also adding more languages as illustrated by the DR02 work should also yield much better performance.

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*,

Condition	N	V	A	R	Global F Measure
SM07	68.7	33.01	65.2	63.1	59.2
RelCont-BL	68.7	35.46	65.2	63.1	59.72
RelCont-Final	69.0	38.66	68.8	69.45	**62.13**
TransCont	67.15	33.88	60.2	68.01	58.35
MERGE: RelCont-BL+TransCont	69.3	36.91	66.7	64.45	60.82
VOTE: RelCont-BL+TransCont	71	37.71	66.5	66.1	61.92
MERGE: RelCont-Final+TransCont	70.7	38.66	69.5	70.45	63.14
VOTE: RelCont-Final+TransCont	74.2	38.26	68.6	71.45	**64.58**

Table 5: F-measure % for all Combined experimental conditions on SV2AW

Condition	N	V	A	R	Global F Measure
SM07	60.9	43.4	57	92.88	53.98
RelCont-BL	60.9	48.5	57	92.88	55.87
RelCont-Final	65	49.2	65.55	92.88	**59.52**
TransCont	53.77	38.43	56.48	88.89	49.29
MERGE: RelCont-BL+TransCont	60.6	49.5	58.85	92.88	56.47
VOTE: RelCont-BL+TransCont	59.3	49.5	59.1	92.88	55.92
MERGE: RelCont-Final+TransCont	63.2	50.3	65.25	92.88	**59.07**
VOTE: RelCont-Final+TransCont	62.4	49.65	65.25	92.88	58.47

Table 6: F-measure % for all Combined experimental conditions on SV3AW

pages 255–262, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Christiane Fellbaum. 1998. "wordnet: An electronic lexical database". MIT Press.

Weiwei Guo and Mona Diab. 2009. Improvements to monolingual english word sense disambiguation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 64–69, Boulder, Colorado, June. Association for Computational Linguistics.

N. Ide and J. Veronis. 1998. Word sense disambiguation: The state of the art. In *Computational Linguistics*, pages 1–40, 24:1.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*. The MIT Press.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *In Proceedings of the SIGDOC Conference*, Toronto, June.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 411–418, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

George A. Miller. 1990. Wordnet: a lexical database for english. In *Communications of the ACM*, pages 39–41.

Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the* 20^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1683–1688, Hyderabad, India.

Roberto Navigli. 2009. Word sense disambiguation: a survey. In *ACM Computing Surveys*, pages 1–69. ACM Press.

Franz Joseph Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, , and H. Dang. 2001. English tasks: all-words and verb lexical sample. In *In Proceedings of ACL/SIGLEX Senseval-2*, Toulouse, France, June.

Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. 2005. Maximizing semantic relatedness to perform word sense disambiguation. In *University of Minnesota Supercomputing Institute Research Report UMSI 2005/25*, Minnesotta, March.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA.

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