Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger*

Massimiliano Ciaramita
Inst. of Cognitive Science and Technology
Italian National Research Council
m.ciaramita@istc.cnr.it

Yasemin Altun
Toyota Technological Institute
at Chicago
altun@tti-c.org

Abstract

In this paper we approach word sense disambiguation and information extraction as a unified tagging problem. The task consists of annotating text with the tagset defined by the 41 Wordnet supersense classes for nouns and verbs. Since the tagset is directly related to Wordnet synsets, the tagger returns partial word sense disambiguation. Furthermore, since the noun tags include the standard named entity detection classes – person, location, organization, time, etc. – the tagger, as a by-product, returns extended named entity information. We cast the problem of supersense tagging as a sequential labeling task and investigate it empirically with a discriminatively-trained Hidden Markov Model. Experimental evaluation on the main sense-annotated datasets available, i.e., Semcor and Senseval, shows considerable improvements over the best known "first-sense" baseline.

1 Introduction

Named entity recognition (NER) is the most studied *information extraction* (IE) task. NER typically focuses on detecting instances of "person", "location", "organization" names and optionally instances of "miscellaneous" or "time" categories. The scalability of statistical NER allowed researchers to apply it successfully on large collections of newswire text, in several languages, and biomedical literature. Newswire NER performance, in terms of F-score, is in the upper 80s (Carreras et al., 2002; Florian et al., 2003), while Bio-NER accuracy ranges between the low 70s and 80s, depending on the data-set used for training/evaluation (Dingare et al., 2005). One shortcoming of NER is its over-simplified ontological model, leaving instances of other potentially informative categories unidentified. Hence, the utility of named entity information is limited. In addition, instances to be detected are mainly restricted to (sequences of) proper nouns.

Word sense disambiguation (WSD) is the task of deciding the intended sense for ambiguous words in context. With respect to NER, WSD lies at the other end of the semantic tagging spectrum, since the dictionary defines tens of thousand of very specific word senses, including NER categories. Wordnet (Fellbaum, 1998)[1], possibly the most used resource for WSD, defines word senses for verbs, common and proper nouns. Word sense disambiguation, at this level of granularity, is a complex task which resisted all attempts of robust broad-coverage solutions. Many distinctions are too subtle to be captured automatically, and the magnitude of the class space – several orders larger than NER's – makes it hard to approach the problem with sophisticated, but scalable, machine learning methods. Lastly, even if the methods would scale up, there are not enough manually tagged data, at the word sense level, for training a model. The performance of state of the art WSD systems on realistic evaluations is only comparable to the "first sense" baseline (cf. Section 5.3). Notwithstanding much research, the benefits of disambiguated lexical information for language processing are still mostly speculative.

This paper presents a novel approach to broad-

*The first author is now at Yahoo! Research. The tagger described in this paper is free software and can be downloaded from http://www.loa-cnr.it/ciaramita.html.

[1] When referring to Wordnet, throughout the paper, we mean Wordnet version 2.0.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 594–602,
Sydney, July 2006. ©2006 Association for Computational Linguistics

NOUNS			
SUPERSENSE	NOUNS DENOTING	SUPERSENSE	NOUNS DENOTING
act	acts or actions	object	natural objects (not man-made)
animal	animals	quantity	quantities and units of measure
artifact	man-made objects	phenomenon	natural phenomena
attribute	attributes of people and objects	plant	plants
body	body parts	possession	possession and transfer of possession
cognition	cognitive processes and contents	process	natural processes
communication	communicative processes and contents	person	people
event	natural events	relation	relations between people or things or ideas
feeling	feelings and emotions	shape	two and three dimensional shapes
food	foods and drinks	state	stable states of affairs
group	groupings of people or objects	substance	substances
location	spatial position	time	time and temporal relations
motive	goals	Tops	abstract terms for unique beginners
VERBS			
SUPERSENSE	VERBS OF	SUPERSENSE	VERBS OF
body	grooming, dressing and bodily care	emotion	feeling
change	size, temperature change, intensifying	motion	walking, flying, swimming
cognition	thinking, judging, analyzing, doubting	perception	seeing, hearing, feeling
communication	telling, asking, ordering, singing	possession	buying, selling, owning
competition	fighting, athletic activities	social	political and social activities and events
consumption	eating and drinking	stative	being, having, spatial relations
contact	touching, hitting, tying, digging	weather	raining, snowing, thawing, thundering
creation	sewing, baking, painting, performing		

Table 1. Nouns and verbs supersense labels, and short description (from the Wordnet documentation).

coverage information extraction and word sense disambiguation. Our goal is to simplify the disambiguation task, for both nouns and verbs, to a level at which it can be approached as any other tagging problem, and can be solved with state of the art methods. As a by-product, this task includes and extends NER. We define a tagset based on Wordnet's lexicographers classes, or *supersenses* (Ciaramita and Johnson, 2003), cf. Table 1. The size of the supersense tagset allows us to adopt a structured learning approach, which takes local dependencies between labels into account. To this extent, we cast the supersense tagging problem as a sequence labeling task and train a discriminative Hidden Markov Model (HMM), based on that of Collins (2002), on the manually annotated Semcor corpus (Miller et al., 1993). In two experiments we evaluate the accuracy of the tagger on the Semcor corpus itself, and on the English "all words" Senseval 3 shared task data (Snyder and Palmer, 2004). The model outperforms remarkably the best known baseline, the first sense heuristic – to the best of our knowledge, for the first time on the most realistic "all words" evaluation setting.

The paper is organized as follows. Section 2 introduces the tagset, Section 3 discusses related work and Section 4 the learning model. Section 5 reports on experimental settings and results. In Section 6 we summarize our contribution and consider directions for further research.

2 Supersense tagset

Wordnet (Fellbaum, 1998) is a broad-coverage machine-readable dictionary which includes 11,306 verbs mapped to 13,508 word senses, called *synsets*, and 114,648 common and proper nouns mapped to 79,689 synsets. Each noun or verb synset is associated with one of 41 broad semantic categories, in order to organize the lexicographer's work of updating and managing the lexicon (see Table 1). Since each lexicographer category groups together many synsets they have been also called *supersenses* (Ciaramita and Johnson, 2003). There are 26 supersenses for nouns, 15 for verbs. This coarse-grained ontology has a number of attractive features, for the purpose of natural language processing. First, the small size of the set makes it possible to build a single tagger which has positive consequences on robustness. Second, classes, although fairly general, are easily recognizable and not too abstract or vague. More importantly, similar word senses tend to be merged together.

As an example, Table 2 summarizes all senses of the noun "box". The 10 synsets are mapped to 6 supersenses: "artifact", "quantity", "shape", "state", "plant", and "act". Three similar senses (2), (7) and (9), and the probably related (8), are merged in the "artifact" supersense. This process can help disambiguation because it removes sub-

1. {box} (container) "he rummaged through a box of spare parts" - n.artifact
2. {box, loge} (private area in a theater or grandstand where a small group can watch the performance) "the royal box was empty" - n.artifact
3. {box, boxful} (the quantity contained in a box) "he gave her a box of chocolates" - n.quantity
4. {corner, box} (a predicament from which a skillful or graceful escape is impossible) "his lying got him into a tight corner" - n.state
5. {box} (a rectangular drawing) "the flowchart contained many boxes" - n.shape
6. {box, boxwood} (evergreen shrubs or small trees) - n.plant
7. {box} (any one of several designated areas on a ball field where the batter or catcher or coaches are positioned) "the umpire warned the batter to stay in the batter's box" - n.artifact
8. {box, box seat} (the driver's seat on a coach) "an armed guard sat in the box with the driver" - n.artifact
9. {box} (separate partitioned area in a public place for a few people) "the sentry stayed in his box to avoid the cold" - n.artifact
10. {box} (a blow with the hand (usually on the ear)) "I gave him a good box on the ear" - n.act

Table 2. The noun "box" in Wordnet: each line lists one synset, the set of synonyms, a definition, an optional example sentence, and the supersense label.

tle distinctions, which are hard to discriminate and increase the size of the class space. One possible drawback is that senses which one might want to keep separate, e.g., the most common sense box/container (1), can be collapsed with others. One might argue that all "artifact" senses share semantic properties which differentiate them from the other senses and can support useful semantic inferences. Unfortunately, there are no general solutions to the problem of sense granularity. However, major senses identified by Wordnet are maintained at the supersense level. Hence, supersense-disambiguated words are also, at least partially, synset-disambiguated.

Since Wordnet includes both proper and common nouns, the new tagset suggests an extended notion of named entity. As well as the usual NER categories, "person", "group", "location", and "time"[2], supersenses include categories such as artifacts, which can be fairly frequent, but usually neglected. To a greater extent than in standard NER, research in Bio-NER has focused on the adoption of richer ontologies for information extraction. Genia (Ohta et al., 2002), for example, is an ontology of 46 classes – with annotated corpus – designed for supporting information extraction in the molecular biology domain. In addition, there is growing interest for extracting *relations* between entities, as a more useful type of IE (cf. (Rosario and Hearst, 2004)).

Supersense tagging is inspired by similar considerations, but in a domain-independent setting; e.g., verb supersenses can label semantic interactions between nominal concepts. The following sentence (Example 1), extracted from the data – further described in Section 5.1 – shows the information captured by the supersense tagset:

(1) *Clara Harris*$_{n.person}$, one of the *guests*$_{n.person}$ in the *box*$_{n.artifact}$, *stood up*$_{v.motion}$ and *demanded*$_{v.communication}$ *water*$_{n.substance}$.

As Example 1 shows there is more information that can be extracted from a sentence than just the names; e.g. the fact that "Clara Harris" and the following "guests" are both tagged as "person" might suggest some sort of co-referentiality, while the coordination of verbs of motion and communication, as in "stood up and demanded", might be useful for language modeling purposes. In such a setting, structured learning methods, e.g., sequential, can help tagging by taking the senses of the neighboring words into account.

3 Related Work

Sequential models are common in NER, POS tagging, shallow parsing, etc.. Most of the work in WSD, instead, has focused on labeling each word individually, possibly revising the assignments of senses at the document level; e.g., following the "one sense per discourse" hypothesis (Gale et al., 1992). Although it seems reasonable to assume that occurrences of word senses in a sentence can be correlated, hence that structured learning methods could be successful, there has not been much work on sequential WSD. Segond et al. (1997) are possibly the first to have applied an HMM tagger to semantic disambiguation. Interestingly, to make the method more tractable, they also used the supersense tagset and estimated the model on Semcor. By cross-validation they show a marked improvement over the first sense baseline. However, in (Segond et al., 1997) the tagset is used differently, by defining equivalence classes of words with the same set of senses. From a similar perspective, de Loupy et al. (de Loupy et al., 1998)

[2] The supersense category "group" is rather a superordinate of "organization" and has wider scope.

also investigated the potential advantages of using HMMs for disambiguation. More recently, variants of the generative HMM have been applied to WSD (Molina et al., 2002; Molina et al., 2004) and evaluated also on Senseval data, showing performance comparable to the first sense baseline.

Previous work on prediction at the supersense level (Ciaramita and Johnson, 2003; Curran, 2005) has focused on lexical acquisition (nouns exclusively), thus aiming at word type classification rather than tagging. As far as applications are concerned, it has been shown that supersense information can support supervised WSD, by providing a partial disambiguation step (Ciaramita et al., 2003). In syntactic parse re-ranking supersenses have been used to build useful latent semantic features (Koo and Collins, 2005). We believe that supersense tagging has the potential to be useful, in combination with other sources of information such as part of speech, domain-specific NER models, chunking or shallow parsing, in tasks such as question answering and information extraction and retrieval, where large amounts of text need to be processed. It is also possible that this kind of shallow semantic information can help building more sophisticated linguistic analysis as in full syntactic parsing and semantic role labeling.

4 Sequence Tagging

We take a sequence labeling approach to learning a model for supersense tagging. Our goal is to learn a function from input vectors, the observations from labeled data, to response variables, the supersense labels. POS tagging, shallow parsing, NP-chunking and NER are all examples of sequence labeling tasks in which performance can be significantly improved by optimizing the choice of labeling over whole sequences of words, rather than individual words. The limitations of the generative approach to sequence tagging, i. e. Hidden Markov Models, have been overcome by discriminative approaches proposed in recent years (McCallum et al., 2000; Lafferty et al., 2001; Collins, 2002; Altun et al., 2003). In this paper we apply perceptron trained HMMs originally proposed in (Collins, 2002).

4.1 Perceptron-trained HMM

HMMs define a probabilistic model for observation/label sequences. The joint model of an observation/label sequence $(\mathbf{x}, \mathbf{y})$, is defined as:

$$P(\mathbf{y}, \mathbf{x}) = \prod_i P(y_i | y_{i-1}) P(x_i | y_i)), \qquad (2)$$

where y_i is the ith label in the sequence and x_i is the ith word. In the NLP literature, a common approach is to model the conditional distribution of label sequences given the label sequences. These models have several advantages over generative models, such as not requiring questionable independence assumptions, optimizing the conditional likelihood directly and employing richer feature representations. This task can be represented as learning a discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, on a training data of observation/label sequences, where F is linear in a feature representation Φ defined over the joint input/output space

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle. \qquad (3)$$

Φ is a global feature representation, mapping each $(\mathbf{x}, \mathbf{y})$ pair to a vector of feature counts $\Phi(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$, where d is the total number of features. This vector is given by

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} \sum_{j=1}^{|\mathbf{y}|} \phi_i(y_{j-1}, y_j, \mathbf{x}). \qquad (4)$$

Each individual feature ϕ_i typically represents a morphological, contextual, or syntactic property, or also the inter-dependence of consecutive labels. These features are described in detail in Section 4.2. Given an observation sequence $\mathbf{x}$, we make a prediction by maximizing F over the response variables:

$$f_{\mathbf{w}}(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}). \qquad (5)$$

This involves computing the Viterbi decoding with respect to the parameter vector $\mathbf{w} \in \mathbb{R}^d$. The complexity of the Viterbi algorithm scales linearly with the length of the sequence.

There are different ways of estimating $\mathbf{w}$ for the described model. We use the perceptron algorithm for sequence tagging (Collins, 2002). The perceptron algorithm focuses on minimizing the error rate, without involving any normalization factors. This property makes it very efficient which is a desirable feature in a task dealing with a large tagset such as ours. Additionally, the performance of perceptron-trained HMMs is very competitive on a number of tasks; e.g., in shallow parsing, where

Algorithm 1 Hidden Markov average perceptron algorithm.

```
Initialize $\mathbf{w}_0 = \vec{0}$
for $t = 1....,T$ do
    Choose $\mathbf{x}^i$
    Compute $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}^i, \mathbf{y}; \mathbf{w})$
    if $\mathbf{y}^i \neq \hat{\mathbf{y}}$ then
        $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \Phi(\mathbf{x}^i, \mathbf{y}^i) - \Phi(\mathbf{x}^i, \hat{\mathbf{y}})$
    end if
    $\mathbf{w} = \frac{1}{T}\sum_t \mathbf{w}_t$
end for
return $\mathbf{w}$
```

the perceptron performance is comparable to that of Conditional Random Field models (Sha and Pereira, 2003), The tendency to overfit of the perceptron can be mitigated in a number of ways including regularization and voting. Here we apply averaging and straightforwardly extended Collins algorithm, summarized in Algorithm 1.

4.2 Features

We used the following combination of spelling/morphological and contextual features. For each observed word x_i in the data ϕ extracts the following features:

1. **Words:** $x_i,\ x_{i-1},\ x_{i-2},\ x_{i+1},\ x_{i+2}$;

2. **First sense:** supersense baseline prediction for x_i, $\mathsf{fs}(x_i)$, cf. Section 5.3;

3. **Combined (1) and (2):** $x_i + \mathsf{fs}(x_i)$;

4. **Pos:** $\mathsf{pos_i}$ (the POS of x_i), $\mathsf{pos_{i-1}}$, $\mathsf{pos_{i-2}}$, $\mathsf{pos_{i+1}}$, $\mathsf{pos_{i+2}}$, $\mathsf{pos_i[0]}$, $\mathsf{pos_{i-1}[0]}$, $\mathsf{pos_{i-2}[0]}$, $\mathsf{pos_{i+1}[0]}$, $\mathsf{pos_{i+2}[0]}$, $\mathsf{pos\_comm_i}$ if x_i's POS tags is "NN" or "NNS" (common nouns), and $\mathsf{pos\_prop_i}$ if x_i's POS is "NNP" or "NNPS" (proper nouns);

5. **Word shape:** $\mathsf{sh}(x_i)$, $\mathsf{sh}(x_{i-1})$, $\mathsf{sh}(x_{i-2})$, $\mathsf{sh}(x_{i+1})$, $\mathsf{sh}(x_{i+2})$, where $\mathsf{sh}(x_i)$ is as described below. In addition $\mathsf{sh_i} = \mathsf{low}$ if the first character of x_i is lowercase, $\mathsf{sh_i} = \mathsf{cap\_brk}$ if the first character of x_i is uppercase and x_{i-1} is a full stop, question or exclamation mark, or x_i is the first word of the sentence, $\mathsf{sh_i} = \mathsf{cap\_nobrk}$ otherwise;

6. **Previous label:** supersense label y_{i-1}.

Word features (1) are morphologically simplified using the morphological functions of the Wordnet library. The first sense feature (2) is the label predicted for x_i by the baseline model, cf. Section 5.3. POS labels (4) were generated using Brants' TnT tagger (Brants, 2002). POS features of the form $\mathsf{pos_i[0]}$ extract the first character from the POS label, thus providing a simplified representation of the POS tag. Finally, word shape features (5) are regular expression-like transformation in which each character c of a string s is substituted with X if c is uppercase, if lowercase, c is substituted with x, if c is a digit it is substituted with d and left as it is otherwise. In addition each sequence of two or more identical characters c is substituted with $c*$. For example, for s = "Merrill Lynch& Co.", $\mathsf{sh(s)} = \mathsf{Xx} * \mathsf{Xx} * \&\mathsf{Xx}..$

Exploratory experiments with richer feature sets, including syntactic information, affixes, and topic labels associated with words, did not result in improvements in terms of performance. While more experiments are needed to investigate the usefulness of other sources of information, the feature set described above, while basic, offers good generalization properties.

5 Experiments

5.1 Data

We experimented with the following data-sets[3]. The Semcor corpus (Miller et al., 1993), a fraction of the Brown corpus (Kučera and Francis, 1967) which has been manually annotated with Wordnet synset labels. Named entities of the categories "person", "location" and "group" are also annotated. The original annotation with Wordnet 1.6 synset IDs has been converted to the most recent version 2.0 of Wordnet. Semcor is divided in three parts: "brown1" and "brown2", here referred to as "SEM", in which nouns, verbs, adjectives and adverbs are annotated. In addition, the section "brownv", "SEMv" here, contains annotations only for verbs. We also experimented with the Senseval-3 English all-words tasks data (Snyder and Palmer, 2004), here called "SE3". The Senseval all-words task evaluates the performance of WSD systems on all open class words in complete documents. The Senseval-3 data consists of two Wall Street Journal Articles, "wsj_1778" and

[3] These datasets are available in a consistent format and can be downloaded from http://www.cs.unt.edu/ rada/downloads.html

	Dataset		
Counts	SE3	SEM	SEMv
Sentences	300	20,138	17,038
Tokens	5,630	434,774	385,546
Supersenses	1,617	135,135	40,911
Verbs	725	47,710	40,911
Nouns	892	87,425	0
Avg-poly-N-WS	4.66	4.41	4.33
Avg-poly-N-SS	2.86	2.75	2.66
Avg-poly-V-WS	11.17	10.87	11.05
Avg-poly-V-SS	4.20	4.11	4.16

Table 3. Statistics of the datasets. The row "Supersenses" lists the number of instances of supersense labels, partitioned, in the following two rows, between verb and noun supersense labels. The lowest four rows summarize average polysemy figures at the synset and supersense level for both nouns and verbs.

"wsj_1695", and a fiction excerpt, "cl_23", from the unannotated portion of the Brown corpus. Table 3 summarizes a few statistics about the composition of the datasets. The four lower rows report the average polysemy of nouns ("N") and verbs ("V"), in each dataset, both at the synset level ("WS") and supersense ("SS") level. The average number of senses decreases significantly when the more general sense inventory is considered.

We substituted the corresponding supersense to each noun and verb synset in all three data-sets: SEM, SEMv and SE3. All other tokens were labeled "0". The supersense label "noun.Tops" refers to 45 synsets which lie at the very top of the Wordnet noun hierarchy. Some of these synsets are expressed by very general nouns such as "biont", "benthos", "whole", and "nothing". However, others undoubtedly refer to other supersenses, for which they provide the label, such as "food", "person", "plant" or "animal". Since these nouns tend to be fairly frequent, it is confusing and inconsistent to label them "noun.Tops"; e.g., nouns such as "chowder" and "Swedish meatball" would be tagged as "noun.food", but the noun "food" would be tagged as "noun.Tops". For this reason, in all obvious cases, we substituted the "noun.Tops" label with the more specific supersense label for the noun[4].

The SEMv dataset only includes supersense labels for verbs. In order to avoid unwanted false negatives, that is, thousands of nouns labeled "0", we applied the following procedure. Rather than using the full sentences from the SEMv dataset, from each sentence we generated the fragments including a verb but no common or proper nouns; e.g., from a sentence such as "Karns' ruling *pertained*$_{verb.stative}$ to eight of the 10 cases." only the fragment "*pertained*$_{verb.stative}$ to eight of the 10" is extracted and used for training.

Sometimes more than one label is assigned to a word, in all data-sets. In these cases we adopted the heuristic of only using the first label in the data as the correct synset/supersense. We leave the extension of the tagger to the multilabel case for future research. As for now, we can expect that this solution will simply lower, somewhat, both the baseline and the tagger performance. Finally, we adopted a beginning (B) and continuation of entity (I) plus no label (0), encoding; i.e., the actual class space defines 83 labels.

5.2 Setup

The supersense tagger was trained on the Semcor datasets SEM and SEMv. The only free parameter to set in evaluation is the number of iterations to perform T (cf. Algorithm 1). We evaluated the model's accuracy on Semcor by splitting the SEM data randomly in training, development and evaluation. In a 5-fold cross-validation setup the tagger was trained on 4/5 of the SEM data, the remaining data was split in two halves, one used to fix T the other for evaluating performance on test. The full SEMv data was always added to the training portion of SEM. We also evaluated the model on the Senseval-3 data, using the same value for T set by cross-validation on the SEM data[5]. The ordering of the training instances is randomized across different runs, therefore the algorithm outputs different results after each run, even if the evaluation set is fixed, as is the case for the Senseval evaluation. The variance in the results on the SE3 data was measured in this way.

5.3 Baseline tagger

The first sense baseline is the supersense of the most frequent synset for a word, according to Wordnet's sense ranking. This baseline is very competitive in WSD tasks, and it is extremely hard to improve upon even slightly. In fact, the baseline has been proposed as a good alternative to WSD

[4] The nouns which are left with the "noun.Top" label are: entity, thing, anything, something, nothing, object, living thing, organism, benthos, heterotroph, life, and biont.

[5] On average T is equal to 12 times the size of the training data.

	Semcor			Senseval-3		
Method	Recall	Precision	F-score [σ]	Recall	Precision	F-score [σ]
Rand	42.99	38.17	40.44	42.09	35.84	38.70
Baseline	69.25	63.90	66.47	68.65	60.10	64.09
Supersense-Tagger	77.71	76.65	77.18 0.45	73.74	67.60	70.54 0.21

Table 4. Summary of results for random and first sense baselines and supersense tagger, σ is the standard error computed on the five trials results.

altogether (cf. (McCarthy et al., 2004)). For this reason we include the first sense prediction as one of the features of our tagging model.

We apply the heuristic as follows. First, in each sentence, we identify the longest sequence which has an entry in Wordnet as either noun or verb. We carry out this step using the Wordnet's library functions, which perform also morphological simplification. Hence, in Example 1 the entry "stand up" is detected, although also "stand" has an entry in Wordnet. Then, each word identified in this way is assigned its most frequent sense – the only one available if the word is unambiguous. To reduce the number of candidate supersenses we distinguish between common and proper nouns; e.g. "Savannah" (city/river) is distinguished from "savannah" (grassland). This method improves slightly the accuracy of the baseline which does not distinguish between different types of nouns.

5.4 Results

Table 4 summarizes overall performance[6]. The first line shows the accuracy of a baseline which assigns possible supersenses of identified words at random. The second line shows the performance of the first sense baseline (cf. Section 5.3), the marked difference between the two is a measure of the robustness of the first sense heuristic. On the Semcor data the tagger improves over the baseline by 10.71%, 31.19% error reduction, while on Senseval-3 the tagger improves over the baseline by 6.45%, 17.96% error reduction. We can put these results in context, although indirectly, by comparison with the results of the Senseval-3 all words task systems. There, with a baseline of 62.40%, only 4 out of 26 systems performed above the baseline, with the two best systems (Mihalcea and Faruque, 2004; Decadt et al., 2004) achieving an F-score of 65.2% (2.8% improvement, 7.45% error reduction). The system based on the HMM tagger (Molina et al., 2004), achieved an F-score of 60.9%. The supersense tagger improves mostly on precision, while also improving on recall. Overall the tagger achieves F-scores between 70.5 and 77.2%. If we compare these figures with the accuracy of NER taggers the results are very encouraging. Given the considerably larger – one order of magnitude – class space some loss has to be expected. Experiments with augmented tagsets in the biomedical domain also show performance loss with respect to smaller tagsets; e.g., Kazama et al. (2002) report an F-score of 56.2% on a tagset of 25 Genia classes, compared to the 75.9% achieved on the simplest binary case. The sequence fragments from SEMv contribute about 1% F-score improvement.

Table 5 focuses on subsets of the evaluation. The upper part summarizes the results on Semcor for the classes comparable to standard NER's: "person", "group", "location" and "time". However, these categories here are composed of common nouns as well as proper names/named entities. On this four tags the tagger achieves an average 82.46% F-score, not too far from NER results. The lower portion of Table 5 summarizes the results on the five most frequent noun and verb supersense labels on the Senseval-3 data, providing more specific evidence for the supersense tagger's disambiguation accuracy. The tagger outperforms the first sense baseline on all categories, with the exception of "verb.cognition" and "noun.person". The latter case has a straightforward explanation, named entities (e.g., "Phil Haney", "Chevron" or "Marina District") are not annotated in the Senseval data, while they are in Semcor. Hence the tagger learns a different model for nouns than the one used to annotate the Senseval data. Because of this discrepancy the tagger tends to return false positives for some categories. In fact, the other noun categories on which the tagger performs poorly in SE3 are "group" and "location" (baseline 52.10 tagger 44.72 and baseline 47.62% tagger 47.54% F-score). Naturally, the lower performance on Senseval is also explained by the fact that the eval-

[6] Scoring was performed with a re-implementation of the "conlleval" script .

NER supersenses in Semcor							
		Supersense-Tagger			Baseline		
Supersense	# Supersenses	R	P	F	R	P	F
n.person	1526	92.04	87.94	**89.94**	56.29	77.35	65.16
n.group	665	75.38	79.56	**77.40**	62.42	66.81	64.54
n.location	459	77.21	75.37	**76.25**	67.88	63.33	65.53
n.time	412	88.36	84.30	**86.27**	78.26	83.88	80.98
5 most frequent verb supersenses in Senseval-3							
Supersense	# Supersenses	R	P	F	R	P	F
v.stative	184	80.33	81.30	**80.81**	72,83	63.81	68.02
v.communication	88	77.53	83.36	**80.33**	71.91	74.42	73.14
v.motion	81	69.63	64.54	**66.98**	58.02	60.26	59.12
v.cognition	61	73.44	67.91	70.56	75.41	71.87	**73.60**
v.change	60	68.33	67.47	**67.89**	56.67	57.63	57.14
5 most frequent noun supersenses in Senseval-3							
Supersense	# Supersenses	R	P	F	R	P	F
n.person	148	92.24	60.49	73.06	89.12	79.39	**83.97**
n.artifact	131	80.91	77.73	**79.29**	74.24	75.97	75.10
n.act	96	61.46	72.37	**66.45**	58.33	65.12	61.54
n.cognition	67	45.80	52.87	**49.06**	49.28	46.58	47.89
n.event	60	70.33	89.83	**78.87**	71.67	75.44	73.50

Table 5. Summary of results of baseline and tagger on selected subsets of labels: NER categories evaluated on Semcor (upper section), and 5 most frequent verb (middle) and noun (bottom) categories evaluated on Senseval.

uation comes from different sources than training.

6 Conclusions

In this paper we presented a novel approach to broad-coverage word sense disambiguation and information extraction. We defined a tagset based on Wordnet supersenses, a much simpler and general semantic model than Wordnet which, however, preserves significant polysemy information and includes standard named entity recognition categories. We showed that in this framework it is possible to perform accurate broad-coverage tagging with state of the art sequence learning methods. The tagger considerably outperformed the most competitive baseline on both Semcor and Senseval data. To the best of our knowledge the results on Senseval data provide the first convincing evidence of the possibility of improving by considerable amounts over the first sense baseline.

We believe both the tagset and the structured learning approach contribute to these results. The simplified representation obviously helps by reducing the number of possible senses for each word (cf. Table 3). Interestingly, the relative improvement in performance is not as large as the relative reduction in polysemy. This indicates that sense granularity is only one of the problems in WSD. More needs to be understood concerning sources of information, and processes, that affect word sense selection in context. As far as the tagger is concerned, we applied the simplest feature representation, more sophisticated features can be used, e.g., based on kernels, which might contribute significantly by allowing complex feature combinations. These results also suggest new directions of research within this model. In particular, the labels occurring in each sequence tend to coincide with predicates (verbs) and arguments (nouns and named entities). A sequential dependency model might not be the most accurate at capturing the grammatical dependencies between these elements. Other conditional models, e.g., designed on head to head, or similar, dependencies could prove more appropriate.

Another interesting issue is the granularity of the tagset. Supersenses seem more practical then synsets for investigating the impact of broad-coverage semantic tagging, but they define a very simplistic ontological model. A natural evolution of this kind of approach might be one which starts by defining a semantic model at an intermediate level of abstraction (cf. (Ciaramita et al., 2005)).

References

Y. Altun, T. Hofmann, and M. Johnson. 2003. Discriminative Learning for Label Sequences via Boosting. In *Proceedings of NIPS 2003*.

T. Brants. 2002. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP 2000*.

X. Carreras, L. Marquez, and L. Padro. 2002. Named Entity Extraction Using AdaBoost. In *Proceedings of CONLL 2002*.

M. Ciaramita and M. Johnson. 2003. Supersense Tagging of Unknown Nouns in WordNet. In *Proceedings of EMNLP 2003*.

M. Ciaramita, T. Hofmann, and M. Johnson. 2003. Hierarchical Semantic Classification: Word Sense Disambiguation with World Knowledge. In *Proceedings of IJCAI 2003*.

M. Ciaramita, S. Sloman, M. Johnson, and E. Upfal. 2005. Hierarchical Preferences in a Broad-Coverage Lexical Taxonomy. In *Proceedings of CogSci 2005*.

M. Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP 2002*, pages 1–8.

J. Curran. 2005. Supersense Tagging of Unknown Nouns Using Semantic Similarity. In *Proceedings of ACL 2005*, pages 26–33.

C. de Loupy, M. El-Beze, and P.F. Marteau. 1998. Word Sense Disambiguation Using HMM Tagger. In *Proceedings of LREC 1998*, pages 1255–1258.

B. Decadt, V. Hoste, W. Daelemans, and A. van der Bosch. 2004. GAMBL, Genetic Algorithm Optimization of Memory-Based WSD. In *Proceedings of SENSEVAL-3/ACL 2004*.

S. Dingare, M. Nissim, J. Finkel, C. Manning, and C. Grover. 2005. A System for Identifying Named Entities in Biomedical Text: How Results from Two Evaluations Reflect on Both the System and the Evaluations. *Comparative and Functional Genomics*, 6:77–85.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named Entity Extraction through Classifier Combination. In *Proceedings of CONLL 2003*.

W. Gale, K. Church, and D. Yarowsky. 1992. One Sense per Discourse. In *Proceedings of the DARPA Workshop on Speech and Natural Language*.

J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain (ACL 2002)*.

T. Koo and M. Collins. 2005. Hidden-Variable Models for Discriminative Reranking. In *Proceedings of EMNLP 2005*.

H. Kučera and W. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*, pages 282–289.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of ICML 2000*, pages 591–598.

D. McCarthy, R. Koeling, and J. Carroll. 2004. Finding Predominant Senses in Untagged Text. In *Proceedings of ACL 2004*.

R. Mihalcea and E. Faruque. 2004. SenseLearner: Minimally Supervised Word Sense Disambiguation for All Words in Open Text. In *Proceedings of SENSEVAL-3/ACL 2004*.

G.A. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A Semantic Concordance. In *Proceedings of the 3 DARPA Workshop on Human Language Technology*, pages 303–308.

A. Molina, F. Pla, and E. Segarra. 2002. A Hidden Markov Model Approach to Word Sense Dsiambiguation. In *Proceedings of IBERAMIA 2002*.

A. Molina, F. Pla, and E. Segarra. 2004. WSD System Based on Specialized Hidden Markov Model (upv-shmm-eaw). In *Proceedings of SENSEVAL-3/ACL 2004*.

Y. Ohta, Y. Tateisi, J. Kim, H. Mima, and J. Tsujii. 2002. The GENIA Corpus: An Annotated Research Abstract Corpus in the Molecular Biology Domain. In *Proceedings of HLT 2002*.

B. Rosario and M. Hearst. 2004. Classifying Semantic Relations in Bioscience Text. In *Proceedings of ACL 2004)*.

F. Segond, A. Schiller, G. Grefenstette, and J.P. Chanod. 1997. An Experiment in Semantic Tagging Using Hidden Markov Model. In *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources (ACL/EACL 1997)*, pages 78–81.

F. Sha and F. Pereira. 2003. Shallow Parsing with Conditional Random Fields. In *Proceedings of HLT-NAACL 2003*, pages 213–220.

B. Snyder and M. Palmer. 2004. The english All-Words Tasks. In *Proceedings of SENSEVAL-3/ACL 2004*.