A Memory-Based Approach for Semantic Role Labeling

Beata Kouchnir
Department of Computational Linguistics
University of Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
kouchnir@sfs.uni-tuebingen.de

1 Introduction

This paper presents a system for Semantic Role Labeling (SRL) for the CoNLL 2004 shared task (Carreras and Màrquez, 2004). The task is divided into two sub-tasks, recognition and labeling. These are performed independently with different feature representations. Both modules are based on the principle of memory-based learning.

For the first module, we use the IOB2 format to determine whether a chunk belongs to an argument or not. Furthermore, we test two different strategies for extracting arguments from the classifier output. The second module labels the extracted arguments with one of the 30 semantic roles.

2 Memory-Based Learning

The concept of Memory-Based Learning (MBL) (Lin and Vitter, 1994) is to classify unseen (test) instances based on their similarity to known (training) instances. In practice, this is done by storing all training in memory without abstraction and computing the similarity between new and old examples based on a distance metric. New instances are then assigned the most frequent class within a set of *k* most similar examples (*k*-nearest neighbors).

Memory-based learning algorithms have proven to be effective for several NLP tasks, including named entity recognition (Hendrickx and van den Bosch, 2003), clause identification (Tjong Kim Sang, 2001) and most relevantly, grammatical relation finding (Buchholz, 2002).

As testing all possible distance metrics in combination with different values for *k* is not feasible, we have limited the experiment to the Overlap and Modified Value Difference (MVDM) metrics. The values for *k* tested each metric were 1, 3, 5, 7, and 9. Even values were omitted in order to avoid ties.

The Overlap metric computes the distance between two instances by adding up the differences between the features. For symbolic features, each mismatch has a value of 1. MVDM, however, allows different degrees of similarity by examining co-occurrence of feature values with target classes. While this concept seems more suitable for the underlying task, it is only reliable when used with large amounts of data. For a more detailed description of the distance metrics, see (Daelemans et al., 2003).

TiMBL[1] (Daelemans et al., 2003), the MBL implementation used in this experiment, is freely available from the ILK research group at Tilburg University.

3 The Recognition Module

This module identifies the arguments of a proposition, without assigning a label. For this task we use the **IOB2** format, where **B** marks an element at the beginning of an argument, **I** an element inside an argument and **O** an element that does not belong to an argument.

As all argument boundaries, except for those within the target verb chunks, coincide with base chunk boundaries, the data is processed by words only within the target verb chunk, and by chunks otherwise.

The recognition module uses the following features:

- **Head word and POS** of the focus element, where the head of a multi-word chunk is its last words.
- **Chunk type**: one of the 12 chunks types, without the B- or I- prefix.
- **Clause information**: whether the element is at the beginning, at the end or inside a clause.
- **Directionality**: whether the focus element comes before the target verb, after the target verb, or coincides with the target verb.
- **Distance**: numerical distance (1 .. *n*) between the focus element and the target verb.

[1] http://ilk.kub.nl/software.html

Metric / *k*	B	I	O
Overlap *k*=1	87.27	69.34	80.49
MVDM *k*=1	85.96	74.22	83.83
Overlap *k*=3	87.91	71.96	82.61
MVDM *k*=3	87.68	75.35	85.12
Overlap *k*=5	88.37	73.43	83.47
MVDM *k*=5	89.21	76.70	86.52
Overlap *k*=7	88.56	73.41	83.54
MVDM *k*=7	89.31	**77.43**	**86.83**
Overlap *k*=9	88.69	73.61	84.04
MVDM *k*=9	**89.39**	77.38	86.77

Table 1: Results for different distance metrics and values of *k*

- **Adjacency**: whether the focus element is adjacent to the verb chunk or not, or it is within the verb chunk.
- The **target verb and voice**: the voice is passive if the target verb is a past participle preceded by a form of *to be*, and active otherwise.
- **Context**: in addition, the features **head word**, **part of speech**, **chunk type** and **adjacency** of the three chunks each to the left and right of the focus chunk are used as context information.

Testing each feature separately showed the directionality and adjacency features to be most useful. Omitting one feature at a time showed to decrease performance for every omitted feature. Therefore, all of the above features were used in the final system.

The best TiMBL parameter setting for this task was determined to be the Modified Value Difference metric paired with a set of seven nearest neighbors. As we anticipated, the nature of the task requires a more subtle differentiation than the Overlap metric can provide. Furthermore, the size of the training set is apparently sufficient to take full advantage of MVDM. The results for both metrics and all values of *k* are summarized in Table 1. It is interesting to observe the effect of the *k* value for each class. Although the results for the I- and O-classes decrease after *k*=7, those for the B-class do not. However, since the overall results are best for *k*=7, this values was chosen for the final system.

For all metric/*k* combination, the results for the I class are much lower than for the other two. The most common error is the assignment of the O class to I-elements, or vice versa. This performance distribution implies that while the beginning of most arguments is recognized correctly, their span is not, which results in many "broken-up" arguments.

To filter out the actual arguments, we try a strict and a lenient approach. For the latter, any sequence of elements that is not labeled as O is considered an argument (i.e. also those not starting with a B-element). Although this approach slightly reduces the number of missed arguments, it also vastly overgenerates, which ultimately decreases performance. The former approach recognizes as arguments only those sequences beginning with a B-element. Since B is the class most reliably predicted by the classifier, this approach yields better overall performance.

4 The Labeling Module

This module assigns one of the 30 semantic role labels to the arguments extracted by the recognition module. Here, we used only ten features, of which four are "recycled" from the previous module:

- **Word, POS and chunk sequence**: the head words of all the chunks in the argument, their respective parts of speech and chunk types. As TiMBL only allows feature vectors of a fixed length, each of the sequences represents one value.
- **Clause information**: as an element sequences can be a whole clause we added this value to the beginning, end and inside values described in Section 3.
- **Length**: the length in chunks of the argument.
- **Directionality and adjacency**: same as in Section 3.
- The **target verb and voice**: same as in Section 3.
- **Prop Bank roleset** of the target verb: as an analysis of the training data showed that about 86% of the verbs were used in their first sense, and many times, the rolesets for the first two senses are identical, we only considered the roleset of first sense.

Just as for the recognition module, the directionality and adjacency features had the highest information gain. The POS sequence and length features showed no effect, and their omission even slightly improved performance. Therefore, the final system uses only eight features.

To test the performance of this module independently from the first, it was evaluated on the gold-standard arguments (i.e. recognition score of 100). While MVDM once again outperforms the Overlap metric, the optimal value for *k* in this setting is one. The former supports the assumption that for feature values such as words, or word sequences, some values are more similar than others. The latter suggest that the size of the nearest neighbor set (1 vs. 7) should be somewhat proportional to the length of the feature vector (8 vs. 45).

The results for each semantic role are summarized in Table 2. It can be seen that arguments with very restricted surface patterns (e.g. AM-DIS, AM-MOD, AM-NEG)

	Precision	Recall	$F_{\beta=1}$
Overall	75.71%	74.60%	75.15
A0	82.35%	83.41%	82.88
A1	80.69%	82.14%	81.40
A2	61.89%	64.68%	63.25
A3	36.18%	36.91%	36.54
A4	58.39%	63.95%	61.04
A5	33.33%	50.00%	40.00
AM	0.00%	0.00%	0.00
AM-ADV	41.89%	35.23%	38.27
AM-CAU	16.67%	9.43%	12.05
AM-DIR	40.91%	30.00%	34.62
AM-DIS	84.04%	87.75%	85.85
AM-EXT	48.72%	38.78%	43.18
AM-LOC	55.06%	42.61%	48.04
AM-MNR	55.81%	35.93%	43.72
AM-MOD	89.90%	96.14%	92.92
AM-NEG	95.52%	97.71%	96.60
AM-PNC	55.32%	26.00%	35.37
AM-PRD	25.00%	33.33%	28.57
AM-REC	0.00%	0.00%	0.00
AM-TMP	70.06%	64.43%	67.12
R-A0	82.63%	85.19%	83.89
R-A1	67.90%	74.32%	70.97
R-A2	72.22%	76.47%	74.29
R-A3	0.00%	0.00%	0.00
R-AM-LOC	100.00%	50.00%	66.67
R-AM-MNR	0.00%	0.00%	0.00
R-AM-TMP	44.44%	66.67%	53.33
V	99.84%	99.84%	99.84

Table 2: Results for the labeling module with perfect argument spans

	Precision	Recall	$F_{\beta=1}$
Overall	44.93%	63.12%	52.50
A0	59.19%	80.31%	68.15
A1	48.03%	63.53%	54.70
A2	24.40%	44.55%	31.53
A3	15.92%	30.87%	21.00
A4	33.06%	55.10%	41.33
A5	25.00%	25.00%	25.00
AM	0.00%	0.00%	0.00
AM-ADV	18.77%	29.55%	22.96
AM-CAU	3.57%	7.55%	4.85
AM-DIR	11.01%	20.00%	14.20
AM-DIS	51.75%	86.76%	64.84
AM-EXT	31.15%	38.78%	34.55
AM-LOC	17.26%	25.22%	20.49
AM-MNR	27.69%	30.84%	29.18
AM-MOD	82.93%	96.14%	89.05
AM-NEG	92.65%	96.18%	94.38
AM-PNC	16.35%	17.00%	16.67
AM-PRD	0.00%	0.00%	0.00
AM-REC	0.00%	0.00%	0.00
AM-TMP	31.09%	47.43%	37.56
R-A0	79.21%	87.04%	82.94
R-A1	54.21%	78.38%	64.09
R-A2	68.42%	76.47%	72.22
R-A3	0.00%	0.00%	0.00
R-AM-LOC	44.44%	100.00%	61.54
R-AM-MNR	0.00%	0.00%	0.00
R-AM-TMP	60.00%	100.00%	75.00
V	98.14%	98.26%	98.20

Table 3: Results for the development set

are fairly easy to predict. However, it must be noted that given the correct span, the complex (and most frequently occurring) arguments A0 and A1 can be also predicted with very high accuracy. On the down side, the accuracy for most adjuncts is rather low, even though their surface patterns are thought to be somewhat restricted (e.g. AM-LOC, AM-TMP, AM-MNR, AM-EXT).

5 Evaluation

Tables 3 and 4 show the final results for the development and test set, respectively. Although each module performs fairly well separately, their combined results are suboptimal. This is probably due to the fact that the labeling module is trained with gold standard arguments, and is not able to deal with noise induced by the recognition module. The argument type whose results suffer the most is A1, because it usually spans over several chunks, and is difficult to retrieve correctly by the recognition module.

Improvements to the system could be made on the syntactic, lexical, as well as semantic levels. Firstly, it is crucial to improve the performance of the recognition module on I-elements. This could either be done by using a head-lexicalized parser, or, on a lower level, by a preprocessing module that resolves prepositional phrase attachment. Performance for adjuncts such as AM-LOC or AM-TMP could be improved, by using gazetteers of trigger words (e.g. Tuesday) or morphemes (e.g. -day). Furthermore, one could use a semantic database such as WordNet to cluster words. Last but not least, more advantage could be taken from the information in Prop Bank, so different representations of the rolesets should be explored.

References

Sabine Buchholz. 2002. *Memory-based grammatical relation finding*. Ph.D. thesis, Tilburg University.

Xavier Carreras and Lluís Màrquez. 2004. Introduction

	Precision	Recall	$F_{\beta=1}$
Overall	56.86%	49.95%	53.18
A0	68.12%	63.05%	65.49
A1	55.79%	53.22%	54.48
A2	30.95%	30.95%	30.95
A3	21.77%	18.00%	19.71
A4	30.56%	44.00%	36.07
A5	0.00%	0.00%	0.00
AA	0.00%	0.00%	0.00
AM-ADV	23.91%	10.75%	14.83
AM-CAU	0.00%	0.00%	0.00
AM-DIR	28.89%	26.00%	27.37
AM-DIS	53.30%	53.05%	53.18
AM-EXT	15.00%	21.43%	17.65
AM-LOC	21.78%	9.65%	13.37
AM-MNR	45.19%	23.92%	31.28
AM-MOD	91.18%	91.99%	91.58
AM-NEG	90.77%	92.91%	91.83
AM-PNC	26.09%	7.06%	11.11
AM-PRD	0.00%	0.00%	0.00
AM-TMP	47.49%	31.73%	38.04
R-A0	82.61%	71.70%	76.77
R-A1	64.91%	52.86%	58.27
R-A2	50.00%	44.44%	47.06
R-A3	0.00%	0.00%	0.00
R-AM-LOC	0.00%	0.00%	0.00
R-AM-MNR	0.00%	0.00%	0.00
R-AM-PNC	0.00%	0.00%	0.00
R-AM-TMP	66.67%	14.29%	23.53
V	97.77%	97.82%	97.79

Table 4: Results for the test set

to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of ConNLL-2004*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. TiMBL: Tilburg memory based learner, version 5.0, reference guide. Technical report, ILK.

Iris Hendrickx and Antal van den Bosch. 2003. Memory-based one-step named-entity recognition: Effects of seed list features, classifier stacking, and unannotated data. In *Proceedings of CoNLL-2003*, pages 176–179.

Jyh-Han Lin and Jeffrey Scott Vitter. 1994. A theory for memory-based learning. *Machine Learning*, 17:1–26.

Erik Tjong Kim Sang. 2001. Memory-based clause identification. In *Proceedings of CoNLL-2001*, pages 67–69.