Deriving an Ambiguous Word's Part-of-Speech Distribution from Unannotated Text

Reinhard Rapp
Universitat Rovira i Virgili
Pl. Imperial Tarraco, 1
E-43005 Tarragona, Spain
reinhard.rapp@urv.cat

Abstract

A distributional method for part-of-speech induction is presented which, in contrast to most previous work, determines the part-of-speech distribution of syntactically ambiguous words without explicitly tagging the underlying text corpus. This is achieved by assuming that the word pair consisting of the left and right neighbor of a particular token is characteristic of the part of speech at this position, and by clustering the neighbor pairs on the basis of their middle words as observed in a large corpus. The results obtained in this way are evaluated by comparing them to the part-of-speech distributions as found in the manually tagged Brown corpus.

1 Introduction

The purpose of this study is to automatically induce a system of word classes that is in agreement with human intuition, and then to assign all possible parts of speech to a given ambiguous or unambiguous word. Two of the pioneering studies concerning this as yet not satisfactorily solved problem are Finch (1993) and Schütze (1993) who classify words according to their context vectors as derived from a corpus. More recent studies try to solve the problem of POS induction by combining distributional and morphological information (Clark, 2003; Freitag, 2004), or by clustering words and projecting them to POS vectors (Rapp, 2005).

Whereas all these studies are based on global co-occurrence vectors who reflect the overall behavior of a word in a corpus, i.e. who in the case of syntactically ambiguous words are based on POS-mixtures, in this paper we raise the question if it is really necessary to use an approach based on mixtures or if there is some way to avoid the mixing beforehand. For this purpose, we suggest to look at local contexts instead of global co-occurrence vectors. As can be seen from human performance, in almost all cases the local context of a syntactically ambiguous word is sufficient to disambiguate its part of speech.

The core assumption underlying our approach, which in the context of cognition and child language has been proposed by Mintz (2003), is that words of a particular part of speech often have the same left and right neighbors, i.e. a pair of such neighbors can be considered to be characteristic of a part of speech. For example, a noun may be surrounded by the pair "*the ... is*", a verb by the pair "*he ... the*", and an adjective by the pair "*the ... thing*". For ease of reference, in the remainder of this paper we call these local contexts *neighbor pairs*. The idea is now to cluster the neighbor pairs on the basis of the middle words they occur with. This way neighbor pairs typical of the same part of speech are grouped together. For classification, a word is assigned to the cluster where its neighbor pairs are found. If its neighbor pairs are spread over several clusters, the word can be assumed to be ambiguous. This way ambiguity detection follows naturally from the methodology.

2 Approach

Let us illustrate our approach by looking at Table 1. The rows in the table are the neighbor pairs that we want to consider, and the columns are suitable middle words as we find them in a corpus. Most words in our example are syntactically unambiguous. Only *link* can be either a noun or a verb and therefore shows the co-occurrence patterns of both. Apart from the particular choice of features, what distinguishes our approach from most others is that we do not cluster the words (columns) which would be the more straightforward thing to do. Instead we cluster the neighbor pairs (rows). Clustering the columns would be fine for unambiguous words, but has the drawback that ambiguous words

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 53–56,
Prague, June 2007.

tend to be assigned only to the cluster relating to their dominant part of speech. This means that no ambiguity detection takes place at this stage.

In contrast, the problem of demixing can be avoided by clustering the rows which leads to the condensed representation as shown in Table 2. The neighbor pairs have been grouped in such a way that the resulting clusters correspond to classes that can be linguistically interpreted as nouns, adjectives, and verbs. As desired, all unambiguous words have been assigned to only a single cluster, and the ambiguous word *link* has been assigned to the two appropriate clusters.

Although it is not obvious from our example, there is a drawback of this approach. The disadvantage is that by avoiding the ambiguity problem for words we introduce it for the neighbor pairs, i.e. ambiguities concerning neighbor pairs are not resolved. Consider, for example, the neighbor pair "*then ... comes*", where the middle word can either be a personal pronoun like *he* or a proper noun like *John*. However, we believe that this is a problem that for several reasons is of less importance: Firstly, we are not explicitly interested in the ambiguities of neighbor pairs. Secondly, the ambiguities of neighbor pairs seem less frequent and less systematic than those of words (an example is the omnipresent noun/verb ambiguity in English), and therefore the risk of misclusterings is lower. Thirdly, this problem can be reduced by considering longer contexts which tend to be less ambiguous. That is, by choosing an appropriate context width a reasonable tradeoff between data sparseness and ambiguity reduction can be chosen.

	car	cup	discuss	link	quick	seek	tall	thin
a ... has	●	●		●				
a ... is	●	●		●				
a ... man					●		●	●
a ... woman					●		●	●
the ... has	●	●		●				
the ... is	●	●		●				
the ... man					●		●	●
the ... woman					●		●	●
to ... a			●	●		●		
to ... the			●	●		●		
you ... a			●	●		●		
you ... the			●	●		●		

Table 1: Matrix of neighbor pairs and their corresponding middle words.

	car	cup	discuss	link	quick	seek	tall	thin
a ... has, a ... is, the ... has, the ... is	●	●		●				
a ... man, a ... woman, the ... man, the ... woman					●		●	●
to ... a, to ... the, you ... a, you ... the			●	●		●		

Table 2: Clusters of neighbor pairs.

3 Implementation

Our computations are based on the 100 million word British National Corpus. As the number of word types and neighbor pairs is prohibitively high in a corpus of this size, we considered only a selected vocabulary, as described in section 4. From all neighbor pairs we chose the top 2000 which had the highest co-occurrence frequency with the union of all words in the vocabulary and did not contain punctuation marks.

By searching through the full corpus, we constructed a matrix as exemplified in Table 1. However, as a large corpus may contain errors and idiosyncrasies, the matrix cells were not filled with binary yes/no decisions, but with the frequency of a word type occurring as the middle word of the respective neighbor pair. Note that we used raw co-occurrence frequencies and did not apply any association measure. However, to account for the large variation in word frequency and to give an equal chance to each word in the subsequent computations, the matrix columns were normalized.

As our method for grouping the rows we used K-means clustering with the cosine coefficient as our similarity measure. The clustering algorithm was started using random initialization. In order to be able to easily compare the clustering results with expectation, the number of clusters was specified to correspond to the number of expected word classes.

After the clustering has been completed, to obtain their centroids, in analogy to Table 2 the column vectors for each cluster are summed up. The centroid values for each word can now be interpreted as evidence of this word belonging to the class described by the respective cluster. For example, if we obtained three clusters corresponding to nouns, verbs, and adjectives, and if the corresponding centroid values for e.g. the word *link* would be 0.7, 0.3, and 0.0, this could be interpreted such that in 70% of its corpus occurrences *link* has the function of a noun, in 30% of the cases it appears as a verb, and that it never occurs as an adjective. Note that the centroid values for a particular word will always add up to 1 since, as mentioned above, the column vectors have been normalized beforehand.

As elaborated in Rapp (2007), another useful application of the centroid vectors is that they allow us to judge the quality of the neighbor pairs with respect to their selectivity regarding a particular word class. If the row vector of a neighbor pair is very similar to the centroid of its cluster, then it can be assumed that this neighbor pair only accepts middle words of the correct class, whereas neighbor pairs with lower similarity to the centroid are probably less selective, i.e. they occasionally allow for words from other clusters.

4 Results

As our test vocabulary we chose a sample of 50 words taken from a previous study (Rapp, 2005). The list of words is included in Table 3 (columns 1 and 8). Columns 2 to 4 and 9 to 11 of Table 3 show the centroid values corresponding to each word after the procedure described in the previous section has been conducted, that is, the 2000 most frequent neighbor pairs of the 50 words were clustered into three groups. For clarity, all values were multiplied by 1000 and rounded.

To facilitate reference, instead of naming each cluster by a number or by specifying the corresponding list of neighbor pairs (as done in Table 2), we manually selected linguistically motivated names, namely *noun*, *verb*, and *adjective*.

If we look at Table 3, we find that some words, such as *encourage*, *imagine*, and *option*, have one value close to 1000, with the other two values in the one digit range. This is a typical pattern for unambiguous words that belong to only one word class. However, perhaps unexpectedly, the majority of words has values in the upper two digit or three digit range in two or even three columns. This means that according to our system most words seem to be ambiguous in one or another way. For example, the word *brief*, although in the majority of cases clearly an adjective in the sense of *short*, can occasionally also occur as a noun (in the sense of *document*) or a verb (in the sense of *to instruct somebody*). In other cases, the occurrences of different parts of speech are more balanced. An example is the verb *to strike* versus the noun *the strike.*

According to our judgment, the results for all words seem roughly plausible. Only the values for *rain* as a noun versus a verb seemed on first glance counterintuitive, but can be explained by the fact that for semantic reasons the verb *rain* usually only occurs in third person singular, i.e. in its inflected form *rains*.

To provide a more objective measure for the quality of the results, columns 5 to 7 and 12 to 14 of Table 3 show the occurrence frequencies of the 50 words as nouns, verbs, and adjectives in the manually POS-tagged Brown corpus, which is probably almost error free (Kuçera, & Francis, 1967). The respective tags in the Brown-tagset are NN, VB, and JJ.

Generally, the POS-distributions of the Brown corpus show a similar pattern as the automatically generated ones. For example, for *drop* the ratios of the automatically generated numbers 334 / 643 / 24 are similar to those of the pattern from the Brown corpus which is 24 / 34 / 1. Overall, for 48 of the 50 words the outcome with regard to the most likely POS is identical, with the two exceptions being the ambiguous words *finance* and *suit.* Although even in these cases the correct two parts of speech obtain the emphasis, the distribution of the weighting among them is somewhat different.

5 Summary and Future Work

A statistical approach has been presented which clusters contextual features (neighbor pairs) as observed in a large text corpus and derives syntactically oriented word classes from the clusters. In addition, for each word a probability of its occurrence as a member of each of the classes is computed.

Of course, many questions are yet to be explored, among them the following: Can a singular value decomposition (to be in effect only temporarily for the purpose of clustering) reduce the problem of data sparseness? Can biclustering (also referred to as co-clustering or two-mode clustering, i.e. the simultaneous clustering of the rows and columns of a matrix) improve results? Does the approach scale to larger vocabularies? Can it be extended to word sense induction by looking at longer distance equivalents to middle words and neighbor pairs (which could be homographs and pairs of words strongly associated to them)? All these are strands of research that we look forward to explore.

	Simulation			Brown Corpus				Simulation			Brown Corpus		
	Noun	Verb	Adj.	NN	VB	JJ		Noun	Verb	Adj.	NN	VB	JJ
accident	978	8	15	33	0	0	lunch	741	198	60	32	1	0
belief	972	17	11	64	0	0	maintain	4	993	3	0	60	0
birth	968	15	18	47	0	0	occur	15	973	13	0	43	0
breath	946	21	33	51	0	0	option	984	10	7	5	0	0
brief	132	50	819	8	0	63	pleasure	931	16	54	60	1	0
broad	59	7	934	0	0	82	protect	4	995	1	0	34	0
busy	22	22	956	0	1	56	prove	5	989	6	0	53	0
catch	71	920	9	3	39	0	quick	47	14	938	1	0	58
critical	51	13	936	0	0	57	rain	881	64	56	66	2	0
cup	957	23	21	43	1	0	reform	756	221	23	23	3	0
dangerous	37	29	934	0	0	46	rural	66	13	921	0	0	46
discuss	3	991	5	0	28	0	screen	842	126	32	42	5	0
drop	334	643	24	24	34	1	seek	8	955	37	0	69	0
drug	944	10	46	20	0	0	serve	20	958	22	0	107	0
empty	48	187	765	0	0	64	slow	43	141	816	0	8	48
encourage	7	990	3	0	46	0	spring	792	130	78	102	6	0
establish	2	995	2	0	58	0	strike	544	424	32	25	22	0
expensive	55	14	931	0	0	44	suit	200	789	11	40	8	0
familiar	42	17	941	0	0	72	surprise	818	141	41	44	5	3
finance	483	473	44	9	18	0	tape	868	109	23	31	0	0
grow	15	973	12	0	61	0	thank	14	983	3	0	35	0
imagine	4	993	4	0	61	0	thin	32	58	912	0	2	90
introduction	989	0	11	28	0	0	tiny	27	1	971	0	0	49
link	667	311	23	12	4	0	wide	9	4	988	0	0	115
lovely	41	7	952	0	0	44	wild	220	6	774	0	0	51

Table 3: List of 50 words and their values (scaled by 1000) from each of the three cluster centroids. For comparison, POS frequencies from the manually tagged Brown corpus are given.

Acknowledgments

This research was supported by a Marie Curie Intra-European Fellowship within the 6th Framework Programme of the European Community.

References

Clark, Alexander (2003). Combining distributional and morphological information for part of speech induction. *Proceedings of 10th EACL Conference*, Budapest, 59–66.

Finch, Steven (1993). *Finding Structure in Language*. PhD Thesis, University of Edinburgh.

Freitag, Dayne (2004). Toward unsupervised whole-corpus tagging. *Proc. of 20th COLING*, Geneva.

Kuçera, Henry; Francis, W. Nelson (1967). *Computational Analysis of Present-Day American English*. Providence, Rhode Island: Brown University Press.

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90, 91–117.

Rapp, Reinhard (2005). A practical solution to the problem of automatic part-of-speech induction from text. *Proceedings of the 43rd ACL Conference, Companion Volume,* Ann Arbor, MI, 77–80.

Rapp, Reinhard (2007). Part-of-speech discovery by clustering contextual features. In: Reinhold Decker and Hans-J. Lenz (eds.): *Advances in Data Analysis. Proceedings of the 30th Conference of the Gesellschaft für Klassifikation.* Heidelberg: Springer, 627–634.

Schütze, Hinrich (1993). Part-of-speech induction from scratch. *Proceedings of the 31st ACL Conference,* Columbus, Ohio, 251–258.