Redefining part-of-speech classes with distributional semantic models

Andrey Kutuzov
Department of Informatics
University of Oslo
andreku@ifi.uio.no

Parts of speech (PoS) are useful abstractions, but still abstractions. Boundaries between them in natural languages are flexible. Sometimes, large open classes of words are situated on the verge between several parts of speech: for example, participles in English are in many respects both verbs and adjectives. In other cases, closed word classes 'intersect', e.g., it is often difficult to tell a determiner from a possessive pronoun. As Houston (1985) puts it, '*Grammatical categories exist along a continuum which does not exhibit sharp boundaries between the categories*'.

When annotating natural language texts for parts of speech, the choice of a PoS tag in many ways depends on the human annotators themselves, but also on the quality of linguistic conventions behind the division into different word classes. That is why there have been several attempts to refine the definitions of parts of speech and to make them more empirically grounded, based on corpora of real texts: see, among others, the seminal work of Biber et al. (1999). The aim of such attempts is to identify clusters of words occurring naturally and corresponding to what we usually call 'parts of speech'. One of the main distance metrics that can be used in detecting such clusters is a distance between distributional features of words (their contexts in a reference training corpus).

We test this approach using predictive models developed in the field of distributional semantics. Recent achievements in training distributional models of language using machine learning allow for robust representations of natural language semantics created in a completely unsupervised way, using only large corpora of raw text. Relations between dense word vectors (embeddings) in the resulting vector space are as a rule used for semantic purposes. But can they be employed to discover something new about grammar and syntax, particularly parts of speech? Do learned embeddings help here? We show that such models do contain a lot of interesting data related to PoS classes.

For several years already it has been known that some information about morphological word classes is indeed stored in distributional models. Words belonging to different parts of speech possess different contexts: in English, articles are typically followed by nouns, verbs are typically accompanied by adverbs and so on. It means that during the training stage, words of one PoS should theoretically cluster together or at least their embeddings should retain some similarity allowing for their separation from words belonging to other parts of speech.

Our hypothesis is that for the majority of words their parts of speech can be inferred from their embeddings in a distributional model. This inference can be considered a classification problem: we are to train an algorithm that takes a word vector as input and outputs its part of speech. If the word embeddings do contain PoS-related data, the properly trained classifier will correctly predict PoS tags for the majority of words: it means that these lexical entities conform to a dominant distributional pattern of their part of speech class. At the same time, the words for which the classifier outputs *incorrect* predictions, are expected to be 'outliers', with distributional patterns different from other words in the same class. These cases are the points of linguistic interest, and in the rest of the paper we mostly concentrate on them.

To test the initial hypothesis, we used the XML Edition of British National Corpus (BNC), a balanced and representative corpus of English language of about 98 million word tokens in size. We produced a version of BNC where all the words were replaced with their lemmas and PoS-tags converted into the Universal Part-of-Speech

Tagset (Petrov et al., 2012). Thus, each token was represented as a concatenation of its lemma and PoS tag (for example, '*love_VERB*' and '*love_NOUN*' yield different word types). We worked with the following 16 Universal tags: **ADJ, ADP, ADV, AUX, CONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, SCONJ, SYM, VERB, X** (punctuation tokens marked with the PUNCT tag were excluded).

Then, a *Continuous Skipgram* embedding model (Mikolov et al., 2013) was trained on this corpus. This model represents the semantics of the words it contains. But at the same time, for each word, a PoS tag is known (from the BNC annotation). It means that is is possible to test how good the word embeddings are in grouping words according to their parts of speech. To this end, we extracted vectors for the 10 000 most frequent words from the resulting model (roughly, these are the words with corpus frequency more than 500). Then, these vectors were used to train a simple logistic regression multinomial classifier aimed to predict the word's part of speech.

The resulting classifier showed a weighted macro-averaged F-score (over all PoS classes) and accuracy equal to 0.98, with 10-fold cross-validation on the training set. This is a significant improvement over the *one-feature* baseline classifier (classify using only one vector dimension with maximum F-value in relation to class tags), with F-score equal to only 0.22. Thus, the results support the hypothesis that word embeddings contain information that allows us to group words together based on their parts of speech. At the same time, we see that this information is not restricted to some particular vector component: rather, it is distributed among several axis of the vector space.

After training the classifier, we were able to use it to detect 'outlying' words in the BNC (judging by the distributional model). So as not to experiment on the same data we had trained our classifier on, we compiled another test set of 17 000 vectors for words with the BNC frequencies between 100 and 500. Compared to the training error reported above we naturally observe a drop in performance when predicting PoS for this unseen data, but the classifier still appears quite robust, yielding an F-score of 0.91.

All in all, 1741 word types (about 10% of the whole test set vocabulary) were classified incorrectly. These are the 'outliers' we were after. We filtered out misclassified word types with '**X**' BNC annotation (they are mostly foreign words or typos). This leaves us with 1558 words for which the classifier assigned part of speech tags different from the ones in the BNC. It probably means that these words' distributional patterns differ somehow from what is more typically observed, and that they tend to exhibit behavior similar to another part of speech. Table 1 shows the most frequent misclassification cases, together accounting for more than 85% of errors.

Table 1. Most frequent PoS misclassifications of the distributional predictor.

#	Actual PoS	Predicted PoS
347	PROPN	NOUN
313	ADJ	NOUN
190	NOUN	ADJ
91	NOUN	PROPN
87	PROPN	ADJ
57	VERB	ADJ
55	NOUN	NUM
52	NUM	NOUN
45	NUM	PROPN
28	ADV	PROPN
25	ADV	NOUN
25	ADJ	PROPN
20	ADV	ADJ

Almost 30% of error types (judging by absolute amount of misclassified words) consist of proper nouns predicted to be common ones and vice versa. These cases do not tell us anything new, as it is obvious that distributionally these two classes of words are very similar, take the same syntactic contexts and hardly can be considered different parts of speech at all.

Another 30% of errors are due to vague boundaries between nominal and adjectival distribution patterns in English: nouns can be modified by both (it seems that cases where a proper noun is mistaken for an adjective are often caused by the same factor). Words like '*materialist_NOUN*', '*starboard_NOUN*' or '*hypertext_NOUN*' are tagged as nouns in the BNC, but they often modify other nouns, and their contexts are so 'adjectival' that the distributional model actually assigned them semantic features highly similar to those of adjectives. Vice versa, '*white-collar_ADJ*' (an adjective in BNC) is regarded as a noun from the point

of view of our model. Indeed, there can be contradicting views on the correct part of speech for this word in phrases like '*and all the other white-collar workers*'. Thus, in this case the distributional model highlights the already known similarity between two word classes.

The cases of verbs mistaken for adjectives seem to be caused mostly by passive participles ('*was overgrown*', '*is indented*', ''), which intuitively are indeed very adjective-like. So, this gives us a set of verbs dominantly (or almost exclusively, like '*to intertwine*' or '*to disillusion*') used in passive. Of course, we will hardly announce such verbs to be adjectives based on that evidence, but at least we can be sure that this sub-class of verbs is clearly semantically and distributionally different from other verbs.

The next numerous type of errors consists of common nouns predicted to be numerals. A quick glance at the data reveals that 90% of these 'nouns' are in fact currency amounts and percentages ('*£70*', '*33%*', '*$1*', etc). It seems reasonable to classify these as numerals, even though they contain some kind of nominative entities inside. Judging by the decisions of the classifier, their contexts do not differ much from those of simple numbers, and their semantics is similar. The Universal Dependencies Treebank is more consistent in this respect: it separates entities like '*1$*' into two tokens: a numeral (NUM) and a symbol (SYM). Consequently, when our classifier was tested on the words from the UD Treebank, there was only one occurrence of this type of error.

Related to this is the inverse case of numerals predicted to be common or proper nouns. The majority of these 'numerals' are years ('*1804*', '*1776*', '*1822*') and decades ('*1820s*', '*60s*' and even '*twelfths*'). Intuitively, such entities do indeed function as nouns ('*I'd like to return to the sixties*'). Anyway, it is difficult to invent a persuasive reason for why '*fifty pounds*' should be tagged as a noun, but '*the year 1776*' as a numeral. So, this points to possible (minor) inconsistencies in the annotation strategy of the BNC. Note that a similar problem exists in the Penn Treebank as well (Manning, 2011).

The cases we described above revealed some inconsistencies in the BNC annotation. However, it seems that with adverbs mistaken for adjectives, we actually found a systematic error in the BNC tagging: these cases are mostly connected to adjectives like '*plain*', '*clear*' or '*sharp*' (including comparative and superlative forms) erroneously tagged in the corpus as adverbs. These cases are not rare: just the three adjectives we mentioned alone appear in the BNC about 600 times with an adverb tag, mostly in clauses of the kind '*the author makes it plain that...*'. Thus, distributional models can actually detect outright errors in PoS-tagged corpora, when incorrectly tagged words strongly tend to cluster with another part of speech. In the UD treebank such examples can also be observed, but they are much fewer and more 'adverbial', like '*it goes **clear** through*'.

To sum up, the analysis of 'boundary cases' detected by a classifier trained on distributional vectors, indeed reveals sub-classes of words lying on the verge between different parts of speech. It also allows for quickly discovering systematic errors or inconsistencies in PoS annotations, whether they be automatic or manual. Thus, discussions about PoS boundaries would benefit from taking this kind of data into consideration. Arguably, word embeddings are good at predicting PoS precisely because part of speech boundaries are not strict, and even sometimes considered to be a non-categorical linguistic phenomenon (Manning, 2015).

References

[Biber et al.1999] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman grammar of spoken and written English*, volume 2. MIT Press.

[Houston1985] Ann Celeste Houston. 1985. *Continuity and change in English morphology: The variable (ING)*. Ph.D. thesis, University of Pennsylvania.

[Manning2011] Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

[Manning2015] Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41:701–707.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Petrov et al.2012] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC 2012*.