Cross-Language Classification of Function Words

Pablo Gamallo*[1], Gabriel P. Lopes[2], and Joaquim F. Da Silva[2]

[1] Departamento de Língua Espanhola
Universidade de Santiago de Compostela, Espanha
pablogam@usc.es
http://www.sintx.usc.es/ gamallo/
[2] CITI, Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa, Portugal
{jfs,gpl}@di.fct.unl.pt

1 Introduction

Our aim is to describe language-independent properties which will be used to classify function words across a wide range of languages. These generic properties are not defined in terms of traditional word distribution, as in most unsupervised methods to learn syntactic categories from raw corpora [2, 5, 6, 1]. Such a distributional information, which is based on Harry's Hypothesis, turns to be inefficient to measure the distance between words across human languages. In contrast, the features of function words we will describe here are generic dimensions that can be easily used to measure the distance between words belonging to different languages. In this paper, we will describe features such as, for instance, the degree of obligatoriness and the degree of dispersion. Regarding obligatoriness, each word is assigned a weight representing its tendency to be omitted or not between two adjacent words [4]. In the scale of obligatoriness, prepositions and conjunctions are likely to be more obligatory than determiners and auxiliary verbs. Regarding to the degree of dispersion, a function word can be assigned a value obtained by computing how sparse are its immediate neighbor words. For instance, determiners appearing before nouns are assigned low values since the nouns they modify are likely to be sparse. By contrast, nouns appearing before prepositions are less sparse because of the syntactic requirements.

So, the main motivation of the article is to discover general principles governing the distribution of function words across a great variety of natural languages. The main objective is to use this generic distribution to learn categories of function words from raw multilingual corpora. The multilingual corpora used in our experiments is constituted by texts written in three languages, English, French, and Portuguese. These texts are are not translations of an original one, i.e., the input of our method is a non-parallel corpus.

.

* Researcher of the program Isidro Parga Pondal, Xunta de Galicia

2 Cross-Language Features of Function Words

We are interested in discovering cross-language statistical features of the more frequent function words, in order to learn and induce their basic categories. Our learning system uses 20 features, but only 1 of them will be briefly outlined in this abstract. Function words are ranked within a scale of obligatoriness. Given a function word w, a measure (OBL) that gives us an idea about how much is w compelled to be part of their contexts, is computed as follows:

$$OBL(w) = \frac{1}{|(x,y) : \exists(x,w,y)|} \sum_{(x,y):\exists(x,w,y)} \frac{f(x,w,y)}{f(x,w,y) + f(x,y)} \tag{1}$$

where $f(x,w,y)$ denotes the frequency of w in a context constituted by the words, x and y, immediately before and after w, and where $|(x,y) : \exists(x,w,y)|$ is the number of unique contexts of w. Moreover, $f(x,y)$ represents the frequency of word x occurring immediately before y. For instance, preposition "of" cannot be omitted from an expression like "president of Ireland", since the bigram "president Ireland" does not occur in the corpus. Most prepositions have a high value within the scale of obligatoriness. On the other hand, demonstrative determiners tend to be optional.

Function words are also provided with information on the degree of sparseness introduced by the words appearing at their right. Following [3], we call this property "right heterogeneity". For a given function word w, this measure (R_HETER) is obtained by computing a mean of conditional probabilities:

$$R\_HETER(w) = \frac{1}{|(y) : \exists(w,y)|} \sum_{(y):\exists(w,y)} \frac{f(w,y)}{f(w)} \tag{2}$$

For instance, in many languages, prepositions are often followed by a limited set of determiners. It means that prepositions are likely to have a quite homogeneous context at their right. At the opposite, determiners can appear before a non restricted set of nouns. Determiners possesses a very heterogeneous right context.

We also consider left heterogeneity. It measures the degree of word sparseness we find to the left of function words:

$$L\_HETER(w) = \frac{1}{|(x) : \exists(x,w)|} \sum_{(x):\exists(x,w)} \frac{f(x,w)}{f(w)} \tag{3}$$

For instance, the past participle "been" has a very homogeneous left context, since it is followed by a very restricted set of verbal forms: "is", "has", "have", ...

Features OBL, R_HETER, and L_HETER are part of the 20 ones we use to characterize the vector space of function words. Similarity between each pair of vectors is computed by the Euclidean distance.

3 Classification of Function Words

We designed a classification algorithm that takes advantage of the knowledge a speaker has of a particular language. The input can be a very reduced list of tagged function words. Let's take an example. In one of our experiments, the start point for classification was the following list of 5 English function words and the corresponding *POS* tags:

PREP with_E

CONJ or_E

ART the_E

DEM this_E

AUX have_E

The tags considered in this experiment were PREP(osition), CONJ(unction), ART(icle), DEM(onstrative), and AUX(iliar). Given this information and a particular multilingual corpus constituted by three languages (English _E, French _F, and Portuguese _P), we obtained the classes plotted in Table 1.

PREP	*with_E* at_E under_E for_E from_E by_E on_E as_E para_P por_P com_P avec_F pour_F sur_F dans_F par_F
CONJ	*or_E* and_E e_P ou_P et_F
ART	*the_E* a_E o_P os_P uma_P um_P as_P la_F le_F les_F l_F
DEM	*this_E* these_E those_E , such_E any _outros_P qual-quer_P este_P esta_P esse_P estas_P outras_P essa_P ces_F cette_F ce_F
AUX	*have_E* has_E is_E are_E será_P são_P pode_P foi_P peut_F est_F a_F

Table 1. Multingual classes of function words.

Then, we split each class in three monolingual word sets. It results in 15 basic classes of function words (see Table 2).

Each basic class depicted in Table 2 will be used to classify other function words within a particular language. For instance, the English classes will allow us to classify rare and ambiguous function words. Such a monolingual classification will follow a standard method based on context distribution.

	ENGLISH
PREP	with, at, under, for, from, by, on, as
CONJ	or, and
ART	the, a
DEM	this, these, such, any
AUX	have, has, is, are
	PORTUGUESE
PREP	para, por, com
CONJ	e, ou
ART	o, os, uma, um, as
DEM	outros, qualquer, esta, este, esse, estas, outras, essa
AUX	será, são, pode, foi
	FRENCH
PREP	avec, pour, sur, dans, par
CONJ	et
ART	la, le, les, l
DEM	ces, cette, ce
AUX	peut, est, a

Table 2. Monolingual classes of function words.

References

1. A Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of 10th European Chapter of the Association for Computational Linguistics*, pages 59–66, 2003.
2. S. Finch and N. Chater. Bootstrapping syntactic categories. In *14th Annual Meeting of the Cogntive Science Society*, pages 820–825, 1992.
3. Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *14th Annual Meeting of the Cogntive Science Society*, pages 173–183, Boston, Massachusettes, 1995.
4. Pablo Gamallo, Joaquim Ferreira da Silva, and Gabriel P. Lopes. A divide-and-conquer approach to acquire syntactic categories. In Yasubumi Sakakibara and Georgios Paliouras, editors, *Proceedings of the ICGI'04 Conference*, Lecture Notes in Computer Science. Berlin: Springer-Verlag., 2004.
5. H. Schutze. Part-of-speech induction from scratch. In *ACL-93*, Ohio State, 1993.
6. T.C. Smith and I.H. Witten. Probability-driven lexical classification: A corpus-based approach. In *PACLING-95*, 1995.