Part II

Regular Papers

Learning Verbal Transitivity Using LogLinear Models*

Nuno Miguel Marques (nmm@di.fct.unl.pt)**[1], Gabriel Pereira Lopes (gpl@di.fct.unl.pt)[1], and Carlos Agra Coelho (coelho@isa.utl.pt)[2]

[1] Dep. Informática - FCT/UNL
[2] Dep. Matemática - ISA/UTL

Abstract. In this paper we show how loglinear models can be used to cluster verbs based on their subcategorization preferences. We describe how the information about the phrases or clauses a verb goes with can be computationally learned from an automatically tagged corpus with 9,333,555 words. We will use loglinear modeling to describe the relation between the acquired counts for the part-of-speech tags co-occurring with the verbs on predetermined positions.Based on these results an unsupervised clustering algorithm will be proposed.

Keywords: Subcategorization Learning from Corpora, Loglinear Modeling, Clustering, Natural Language Processing.

1 Introduction

Every word in every language has preferences about the phrases it may combine with. The set of syntactic restrictions a word imposes on its arguments (the phrases or clauses that follow that word) is called word syntactic subcategorization. This paper describes work done in order to automatically extract Portuguese verbal subcategorization from an automatically tagged Portuguese corpus ([ML96]) with $9,333,555$ words.

Brent [Bre93] proposed an approach where each subcategorization frame could be extracted by using a small set of highly specific and discriminating morpho-syntactic cues (ex. pronoun *me*). A probabilistic filter, based on a binomial assumption about the presence or absence of each cue, was used to determine if each cue co-occurs with a verb by chance, or, otherwise, if it signals a given subcategorization frame. Only highly accurate, but extremely rare, cues were used. More recently, Manning [Man93] and Briscoe and Carroll [BC97] replaced Brent's cues by using a part-of-speech tagger and a grammar (a simple finite state grammar by Manning and a wide coverage partial parser by Briscoe and Carroll). In order to overcome problems due to verbs with unusual patterns

* Work supported by JNICT Projects CORPUS (PLUS/C/LIN/805/93) and DIXIT (2/2.1/TIT/1670/95)

** Work supported by PhD scholarship JNICT-PRAXIS XXI/BD/2909/94

Ushioda et all. [UEGW96] present an extension to the previous work. Some verbs were manually assigned their correct subcategorization frames. Based on this manually assigned information and on the counts acquired by using regular expression grammar rules, loglinear supervised statistical learning [Fra96] was used to correct the assigned subcategorization classes.

In this paper it will be shown how the independence loglinear model ([Agr90]) can be successfully used for clustering verbs with the same subcategorization behavior. In next section loglinear independence model will be introduced. Later, some experiments will be described in order to pave the ground for introducing a novel clustering algorithm based on loglinear modeling. According to an evaluation presented in [MLC98] this algorithm seems to accurately distinguish between sets of verbs with the same subcategorization behavior.

2 Modeling Transitivity

The Independence Loglinear Model We trained a neural network based part-of-speech tagger [ML96] using a 5000 word manually tagged corpus. This trained part-of-speech tagger was used for automatically tagging text from LUSA corpus with $9,333,555$ words. Then we counted the number of times each part of speech occurred with each verb in the tagged corpus taking into account positional information, i.e. we counted each tag in positions ranging from -5 (five words to the left of our verb) up to +5 (five words to the right of our verb). For the purposes of this paper we will only use global frequencies of articles in position 1 (just after the verb) and nouns in position 2 (that is the total count of *verb, any tag, noun*). Obtained counts are shown in the left part of table 1.

Tables such as this one are usually called contingency tables. We will try to model the relationship between a row variable X (associated to different verbs) and a column variable Y (associated to the part-of-speech co-occurring with the row verb). We assume an independent Poisson model for the frequencies in that table, since our tagger works on a word by word basis, taking only one word as contextual information. In that sense it is different from other known taggers where the best tagging is chosen according to a Viterbi searching procedure over an Hidden Markov Model. So we think the independent Poisson sample assumption still holds.When the two variables we are modeling are independent, we can use the independence model [Agr90] :

$$logE_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \qquad (i = 1, ..., I; j = 1, ..., J).$$

where $logE_{ij}$ is the logarithm of the expected frequency of cell (i, j) and equals the sum of a constant λ with a row parameter λ_i^X and a column parameter λ_j^Y.

Maximum likelihood techniques have been developed for estimating expected frequencies for loglinear models. The `GLIM` package (Numerical Algorithms Group 1986, [Hea88]) was used to fit our parameters. `GLIM` uses the Newton-Raphson method for ML model fitting.

The main advantage of this approach over previous ones is that we have a way to measure if our data is really independent. This is done by comparing the estimated values with the real ones. GLIM presents us with the likelihood-ratio statistics:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}\log(\frac{O_{ij}}{E_{ij}})$$

where O_{ij} are the observed frequencies for cell (i, j). When a model holds, this statistic has a large-sample chi-squared distribution with $(I-1)(J-1)$ degrees of freedom[Agr90].

Verbal Behavior By using loglinear models we are able to analyze the interactions among several part-of-speech tags (either for the same verb, or for two or more different verbs). By not rejecting the independence assumption we are both saying that the row ordering is irrelevant, and that the pattern of evolution of the values in each row, that is the part-of-speech information, may be used to help estimating the values in other rows. This way, by joining several verbs together we are also clustering verbs with similar part-of-speech preferences, and probably with the same subcategorization behavior.

Without loss of generality, we performed a first study on noun phrase subcategorization. Since, the most common noun phrase pattern is built by an article followed by a noun, we have built contingency tables that describe positions 1 and 2 for articles (Art_1) and nouns (N_2)[1], respectively. Then we selected as transitive verbs Portuguese verbs *abandonar* (*to abandon*), *integrar* (*to integrate*), *prometer* (*to promise*), *provocar* (*to provoke*). Verb *cair* (*to fall*), was used as a negative example (as an intransitive verb). Acting like this we aimed at having a clear perspective about how good were the used tags and their position for modeling noun phrase subcategorization. We have compared the column variable $X =< Art_1, N_2 >$ with the row variable Y containing the verbs under study. Left part of table 1 presents the observed frequencies for these verbs. Right part of table 1 presents, in the lower diagonal matrix, the scaled deviance returned by the independence model for the verb pairs we have experimented. The upper diagonal matrix displays the scaled deviance for groups containing more than two verbs. Subscripts give the number of degrees of freedom for each model. If a significance rate of 95% is used, then all models marked with a '*' should be rejected.

Results obtained, as we will see ahead, seem to confirm that the pair $< article_1 - noun_2 >$, relates verbs with the same transitive behavior. Indeed, by comparing all the possible pairs of transitive verbs, only to the pair (*to provoke,to integrate*) does not fit a loglinear independence model. After analysis of our data, we have found that indeed this verb has a different pattern. Although it subcategorizes a noun phrase, the structure of this noun phrase is different from the usual. In the case of the verb *to provoke*, articles are immediately followed by possessive pronouns, more frequently than in the other cases. This

[1] We will assume that subscripts in the tag names are referring to the tag position to the right (or left, if the index is negative) of the verb.

	Art_1	N_2
to abandon	374	272
to integrate	494	332
to provoke	425	364
to promise	94	82
to fall	29	54

	abandon	integrate	promise	provoke	fall
abandon	-	2.5291_2		6.9401_3	$*20.095_3$
integrate	$0,54717_1$	-			
promise	$1,1304_1$	$2,4288_1$	-		
provoke	$2,3387_1$	$*5,8095_1$	$0,012064_1$	-	$*22,612_4$
fall	$*15,673_1$	$*18.901_1$	$*7.8144_1$	$*10.882_1$	-

Table 1. The left table displays observed frequencies for articles and nouns in positions +1 and +2, respectively. The studied verbs are represented in column 1. The right table displays the residuals for fitted loglinear models at the 95% significance level. Subscripts represent degrees of freedom

led us to conclude that in terms of the noun phrases it accepts, this verb selects a different syntactic pattern. Currently used subcategorization methodologies ([Bre93], [Man93], [BC97],[UEGW96]) aren't able to distinguish this peculiar preference from the other preferences. We also noticed that verb *cair* (to fall) doesn't fit with any other considered verb. This is the intended behavior since none of the other verbs is also intransitive.

We seem to have found a good way to measure the degree of association between verbs, taken either individually or in group. These observations led us to propose a methodology, described in the next section, for automatically finding verbs with similar subcategorization patterns.

3 A Clustering Algorithm

If we have a group of verbs $\overrightarrow{v_1}$ and a candidate verb v_2, by modeling the contingency table $X =< Art_1, N_2 >$, $Y =< \overrightarrow{v_1}, v_2 >$, we will be able to decide if verb v_2 has the same transitive behavior as the group of verbs $\overrightarrow{v_1}$. We have used this property to build the following, very simple, clustering algorithm:

1. We start with a list of N verbs $V =< v_1, ..., v_N >$, occurring in a Corpus C, having for each verb v_i their frequency vector X_i (e.g. $X_i =< freq(Art_1), freq(N_2) >_i$).
2. Verbs in V are sorted by decreasing order of the sum of their feature values (e.g. $freq(Art_1)_i + freq(N_2)_i$).
3. set List-of-clusters equal to a single cluster containing a single verb, the most frequent verb.
4. For each v_i in V do
   (a) Join v_i to the group $\overrightarrow{v_j}$ in $List-of-clusters$ where the independence model best explains the contingency table for $Y \times X$ (e.g. the table $Y =< \overrightarrow{v_j}, v_i >$, $X =< Art_1, N_2 >$).
   (b) If v_i doesn't fit with any of the models in $List-of-clusters$ add a new cluster to the list containing v_i.

Verbs are analyzed by order of frequency, this allows us to use first the most informative verbs to define our seed clusters. In the end, we will get K verb clusters, and K loglinear, independent models. This means that we are able to calculate expected frequencies of modeled tags and positions for every verb in a cluster within a certain confidence interval.

We want groups of verbs to be homogeneous in what concerns the distributions of frequencies of article at position 1 and noun at position 2. In order to accomplish this goal, in each step, verbs were added to the cluster only if a loglinear model of independence still fits to the table where this verb was added. The verb *to provoke* illustrates a possible problem with this approach. If we have a verb (such as *to provoke*), that individually doesn't fit with some verbs in a low deviance accepted cluster (such as the cluster of verbs *to abandon*, *to integrate* and *to promise*), and add it to that cluster, then, an independence model can be fitted with the new set of verbs. A possible solution to this could involve taking into account the increase in scaled deviance. Adding the verb *to provoke* to the previously referred cluster shouldn't be allowed since it results in an increase in the scaled deviance (4.411) that is greater than the 95^{th} chi-square percentile for the number of degrees of freedom added, 1, that is 3.841455. This change could result in an undesired higher number of clusters for our algorithm, and it was not implemented in current version.

Generalization of the proposed algorithm has been experimented [MLC98] but due to space restrictions, we can not fully describe those experiments in this paper. However we will explain our main claims. There are part-of-speech tags that can be used for flagging out the presence of certain subcategorization frames. We could use infinitive verb tag for concluding about the existence of an infinitive clause; article tag for a noun phrase; preposition tag for prepositional phrases and subordinated conjunction tag for subordinated clauses.

As we are also interested on the evaluation of our clustering algorithm, in order to have a rough estimate of how well our method is performing, we extracted a list of verbs from our corpus and tagged them as transitive or intransitive, according to Porto Editora's Portuguese dictionary classification. Clusters have been defined as transitive or intransitive based on the classification in the dictionary of the most frequent verb in that cluster. Until now we have found that the best results are achieved when we use a tag that signals the presence of the phrase we want to detect and its complement (the frequency of the verb minus the frequency of the tag). Using a sample of 81 Portuguese verbs, obtained results (over our corpus), have 96.64% precision and 99.17% recall for noun phrases, if evaluated on a corpus where the precision baseline (tagging all verbs as transitive) is 91.35%[2]. Evaluation on prepositional phrases headed by Portuguese preposition *a* (*to*) achieved 92.38% precision and 100.0% recall, for a 52,86% precision baseline. We discuss these results in detail elsewhere ([MLC98]).

[2] As usual, precision is the percentage of correctly tagged verbs (correctly tagged verbs/total verbs) and recall is the percentage of tagged verbs that were correctly tagged (correctly tagged verbs/total of verbs tagged).

4 Conclusions

In our work we have found that the linguistic restrictions imposed by a word on its arguments are more informative than we could expect. Ushioda et alia and Briscoe and Carroll ([BC97], [UEGW96]) report that several subcategorization frames are recognized for each verb and different ranks are assigned to those frames. Yet we have found that even for a simple subcategorization frame, such as the noun phrase direct object for verbs, several types of distinct verbal preferences were observed *inside* the subcategorization frame. If we ever wanted to uncover this kind of phenomena, we should never use a limited predefined set of pattern preferences. By expressing subcategorization as model parameters for contingency tables we can really learn from data.

In the described methodology, no grammar knowledge was required by the system. Correlations were found between words taking into account occurrence frequencies of the distinct part-of-speech tags following analyzed verbs. It is now a task on grammar development the use of the provided lexical subcategorization information, by incorporating it in the lexicon and by adapting existing grammar rules in order to use this quantified lexical information. This approach has the advantage of not committing the extraction process with a particular grammar formalism. The data provided by our analysis once it is incorporated in a lexicon can increase efficiency at several parsing levels.

References

[Agr90] Alan Agresti. *Categorical Data Analysis.* John Wiley and Sons, 1990.

[BC97] Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, 1997.

[Bre93] Michael R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computacional Linguistics*, 19(2):245–262, 1993.

[Fra96] Alexander Franz. *Automatic Ambiguity Resolution in Natural Language Processing*, volume 1171 of *Lecture Notes in Artificial Intelligence.* Springer, 1996.

[Hea88] M. J. R. Healy. *GLIM: An Introduction.* Clarendon Press, Oxford, 1988.

[Man93] Cristopher Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of ACL*, pages 235–242, 1993.

[ML96] Nuno C. Marques and José Gabriel Lopes. A neural network approach to part-of-speech tagging. In *Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese*, pages 1–9, Curitiba, Brazil, October 21-22 1996.

[MLC98] Nuno Miguel Cavalheiro Marques, Gabriel Pereira Lopes, and Carlos Agra Coelho. Using loglinear clustering for subcategorization identification. In *Coling-ACL, submitted paper*, Available at http:\\www-ssdi.di.fct.unl.pt\ ∼nmm, 1998.

[UEGW96] Akira Ushioda, David Evans, Ted Gibson, and Alex Waibel. Estimation of verb subcategorization frame frequencies based on syntactic and multidimensional statistical analysis. In Harry Bunt and Masaru Tomita, editors, *Recent Advances in Parsing Technology.* Kluwer Academic Publishers, 1996.