Automatic Extraction of Polish Verb Subcategorization
An Evaluation of Common Statistics

Jakub Fast and Adam Przepiórkowski

Collegium Invisibile, ul. Krakowskie Przedmieście 3 pok. 12, Warsaw, Poland
and
Institute of Computer Science, Polish Academy of Sciences, ul. Ordona 21, Warsaw, Poland

kuba.fast@gmail.com
adamp@ipipan.waw.pl

Abstract

This article compares and evaluates common statistics used in the process of filtering the hypotheses within the task of automatic valence extraction. A broader range of statistics is compared than the ones usually found in the literature, including Binomial Miscue Probability, Likelihood Ratio, *t* Test, and various simpler statistics. All experiments are performed on the basis of morphosyntactically annotated but very noisy Polish data. Despite a different experimental methodology, the results confirm Korhonen's findings that statistics based solely on the number of occurrences of a given verb and the number of cooccurrences of the verb and a given frame in general fare much better than statistics comparing such conditional frame frequency with the unconditional frame frequency.

1. Introduction

Valence dictionaries are crucial resources in Natural Language Processing, and yet, for many languages such resources are unavailable or they are available in paper form only. Early 1990s saw the advent of the use of corpora and statistical methods for the automatic learning of valence information, but it has been noted in the literature (cf., e.g., (Korhonen, 2002)) that some of the commonly used statistics are less appropriate for the task at hand.

The aim of this paper is to evaluate such common statistics, as applied to very noisy data: in the experiments reported below, linguistic cues are identified by a simple and error-prone shallow grammar on the basis of a corpus automatically annotated with the help of a preliminary version of a morphological analyzer and a statistical disambiguator with a rather high 9.4% error rate (Dębowski, 2004).

The rest of the paper is structured as follows. §2. briefly describes the linguistic input to the statistical module, while the next section, §3., introduces the statistics employed in the experiments. The following section, §4., describes the setup of the experiments. Finally, §5. discusses the results, while §6. compares them to the results of similar experiments reported in the literature.

2. Linguistic Data

The textual material for the experiments reported in this paper is the IPI PAN Corpus of Polish (Przepiórkowski, 2004), the first and currently the only large publicly available morphosyntactically annotated corpus of Polish (cf. `www.korpus.pl`). Since the corpus is rather large (over 300 million segments), its 15-million segment (over 12 million orthographic words; punctuation marks and, in some special cases, clitic-like elements are treated as separate segments) subcorpus, `sample`, was used in the experiments. The corpus does not contain any constituent annotation apart from sentence boundary markers, but it employs a detailed positional tagset providing information about parts of speech, as well as values of inflectional and morphosyntactic categories (Przepiórkowski and Woliński, 2003).

The process of collecting valence cues consists of four steps. First, a simple shallow grammar is applied to the XML corpus sources, resulting in the identification of some NPs, PPs and verbs. Second, each sentence is split into clauses, on the basis of those punctuation marks and conjunctions which are not constituents of NPs and PPs. Third, for each clause containing exactly one verb, V, all NPs and PPs identified in this clause are collected into an observed frame F, and the pair $\langle V, F\rangle$ is added to the set of observations. Finally, all observations are collected into hypotheses represented by tuples $\langle\langle V, F\rangle, n, f, k\rangle$, where $\langle V, F\rangle$ is a verb/frame combination, n is the number of the verb's occurrences in the cue set, f is the total number of the occurrences of the frame, and k is the number of clauses in which they cooccur.

A simple cascade of regular grammars with some added unification-like functionality is used for the shallow parsing of the input and for handling NP- and PP-internal agreement. The whole grammar consists of 18 rules and, consequently, the range of phrases identified by the grammar is very limited: numeral phrases, adjectival phrases, adverbial phrases, clauses and infinitival verbal phrases are excluded from consideration here, i.e., the task at hand is constrained to the identification of possible NP and PP arguments. Two important simplifications in the grammar concern the treatment of nominative and genitive NPs: the former are ignored altogether, i.e., no attempt at distinguishing subject-taking verbs and subjectless verbs is made, while the latter are attached to the immediately preceding NPs and PPs whenever possible, rather than being always treated as potential arguments of verbs.

3. Statistics

Once all the hypotheses are collected, they are rated depending on the dependability of the evidence they provide for inferring that a given frame is valid for a given verb. Two classes of statistics were used for evaluating the strength of the hypotheses: the first class, discussed in §3.2., is composed of metrics which exclude certain hypotheses due to an insufficient verb/frame cooccurrence count given the number of verb's occurrences attested in the cue set; and the second class (§3.3.) judges a given frame as likely to be valid for a given verb if the verb's statistical association with the frame is higher than average for all other verbs in the cue set.

3.1. Probabilistic Model

The statistics presented below share a common probabilistic model. The probability of a frame F occurring given a verb V is taken to be Bernoulli-distributed, i.e., the event-space is defined as that of a single weighted coin toss, where success is defined as an occurrence of F, and failure as the occurrence of some other frame. This model is represented by a random variable $X_1 \sim \text{Be}(\pi_1)$, where π_1 is the theoretical, conditional probability of F occurring in a clause that contains V. A complementary random variable $X_2 \sim \text{Be}(\pi_2)$ will also be taken into consideration in the model, representing the probability of F occurring given some verb *other than* V.

On the basis of this model, and given a number C equal to the total number of clauses in the cue set, a hypothesis of the form $\langle\langle V, F\rangle, n_1, f, k_1\rangle$ is interpreted as describing two samples m_1 and m_2 taken from X_1 and X_2, respectively. m_1's size is taken to correspond to the number of V's occurrences, n_1, and the number of positive outcomes in m_1 is the number of F's occurrences with V, i.e., k_1. The size of m_2 is equal to the total number of clauses that do not contain V ($n_2 = C - n_1$), and the number of successes corresponds to the number of F's occurrences with verbs other than V ($k_2 = f - k_1$).

The elements of each sample are assumed to be independent. For i random variables $Y_1, Y_2, \cdots, Y_i$ with an identical distribution $\text{Be}(\pi)$, the sum $Y = \sum_{j=1}^{i} Y_j$ (i.e., a random variable representing the total number of successes in a sample drawn from those i variables) has a binomial distribution $\text{Bin}(i, \pi)$. Thus, the probability of F occurring k_1 times given n_1 occurrences of V is represented by the random variable $M_1 \sim \text{Bin}(n_1, \pi_1)$, and the probability of F occurring k_2 times given n_2 occurrences of some other verb is represented by the random variable $M_2 \sim \text{Bin}(n_2, \pi_2)$. Everywhere in this text, π_1 and π_2 are estimated on the basis of m_1 and m_2 as, respectively, $\hat{\pi}_1 = p_1 = \frac{k_1}{n_1}$ and $\hat{\pi}_2 = p_2 = \frac{k_2}{n_2}$.

3.2. Minimum Significant Count Statistics

Minimum Significant Count (MSC) statistics rate a given hypothesis on the basis of the numeric relation between k_1 and n_1, assigning every k_1, n_1 some measure of how likely it is for k_1 occurrences of F to have been observed in n_1 trials because of noise.

The general form of an MSC is $S = \phi(k_1, n_1)$ where ϕ is any function monotonically increasing with k_1 for a set n_1. A given F is considered a valid frame for V is made if S exceeds a certain critical value c, below which the evaluated cooccurrence count is deemed accidental.

3.2.1. Binomial Miscue Probability

For a certain independently established probability B_F that a frame F occurs with a verb V even though it is not a valid frame for this verb, the Binomial Miscue Probability (BMP) is the probability of k_1 or more occurrences of F in n_1 trials being produced by a 'noise-generating' random variable $Z \sim \text{Bin}(n_1, B_F)$. BMP was first introduced in (Brent, 1993).[1] The formula for BMP is the following:

$$\text{BMP}_{B_F}(k_1, n_1) = 1 - \Phi_Z(k_1) \quad (1)$$

where Φ_Z is the distribution function for Z. Note that in the case of the formula above, the smaller the value of BMP, the more likely it is for F to be a valid frame for V.

3.2.2. Baseline: Relative Frame Frequency

The baseline MSC consists simply of taking the relative frame frequency for a given verb ($p_1 = \frac{k_1}{n_1}$), and rejecting those verb/frame combinations, for which the resultant value is lower than some threshold.[2]

3.3. Strength of Association Statistics

The Strength of Association (SOA) statistics are based, roughly, on comparing the conditional and unconditional distributions of a given frame by assessing the significance of the difference between p_1 and p_2. The expectation is that if p_1 is significantly lower than p_2, i.e., F occurs with V much less often than it does otherwise, F should be classified as an invalid frame for V.

3.3.1. Likelihood Ratio

The Likelihood Ratio LR statistic is based on comparing the probability that m_1 and m_2 were generated by the best among the models stipulating that $\pi_1 = \pi_2$ and the best one among those that do not need to satisfy this condition. Given that the best fit for the latter model is given by a joint distribution of M_1 and M_2, for $B_{n,p} \sim \text{Bin}(n, p)$, a value λ is calculated with the following formula:

$$\lambda = \frac{\max_p P(B_{n_1,p} = k_1, B_{n_2,p} = k_2\rangle}{P\,(M_1 = k_1, M_2 = k_2)} \quad (2)$$

where the maximal value of p equals $\frac{k_1+k_2}{n_1+n_2}$ and the relevant probabilities are calculated straightforwardly from the aproppriate probability density functions for the binomial distribution. Low values of lambda imply that the two models are distinct, i.e., the values of π_1 and π_2 differ significantly. An assymetrical LR statistic is given by

$$\text{LR}_{\pm} = -2 \log \lambda \times b \quad (3)$$

where $b = -1$ if $p_1 - p_2 < 0$ and 1 otherwise. The resultant statistic has a distribution related to $\chi^2_{df=1}$, and b is introduced in order to distinguish between V strongly favoring F and strongly disfavoring it. If the value of LR is

[1] BMP is also referred to as the Binomial Hypothesis Test.

[2] In the literature on valence extraction, this particular statistic is referred to as the Maximum Likelihood Estimate (of π_1).

lower than a certain critical value, F occurs with V significantly less than with other verbs, and should be classified as an invalid frame for V.

3.3.2. *t* Test

The *t* test measures the significance of the difference between the means of two independent samples. The formula for *t* is the following:

$$t = \frac{p_1 - p_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (4)$$

where s_i^2 is the estimate of the variance of X_i, calculated as $p_i(1-p_i)$. Like with LR, low values of *t* mean that p_1 is significantly lower than p_2, and therefore F is not a valid frame for V. The distribution of *t* is $N(0,1)$.

3.3.3. Baseline: Probability Ratio and Difference of Probabilty

The two statistics above were matched against two trivial measures of the difference between p_1 and p_2: $\frac{p_1}{p_2}$ and $p_1 - p_2$. In both these cases, the expectation was that the lower the value of such statistic, the less likely it is for F to be a valid frame for V.

4. Experiments

The performance of the statistics was evaluated in four experiments, the results of which are presented in Table 1. First, the shallow parsing mechanism described in §2. was applied to four distinct cue sets: one consisting of hypotheses containing all attested $\langle V, F\rangle$ combinations (ALL), and three cue sets containing only hypotheses concerning frames within a given frame frequency range: high (HF, with $f \geq 0.01 \times C$), average (MF, $0.001 \times C \leq f < 0.01 \times C$), and low (LF, $f < 0.001 \times C$). The values for all six statistics were then calculated for the four cue sets and matched against a baseline treatment consisting in considering all the frames seen with a given verb as valid.

The gold standard adopted for the purpose of evaluating the results of these experiments was Marek Świdziński's machine readable valence dictionary (Świdziński, 1998) containing 1492 entries.[3] This dictionary was processed by conflating multiple entries with the same lemma to single entries, which reduced the dictionary to 1369 entries, by translating the original sometimes complex notation, which allowed for optionality and disjunction, into sequences of atomic valence frames, and by removing all frames containing specifications different than NPs and PPs. From the resulting dictionary, 100 frequent verbs (each occurring at least 100 times in the cue set) evenly distributed across the scale of the number of occurrences were blindly selected as the training set (see below), and other 100 verbs, not necessarily frequent, were selected the same way as the evaluation set. The evaluation set was then tailored to each specific cue set by removing frames which do not fall within the particular frame frequency category, therefore the recall values for HF, MF, and LF are calculated in relation to a standard containing only high, average, and low frequency frames, respectively.

The critical (cutoff) values for each statistic were established experimentally. For each cue set and each statistic, an exhaustive search was performed through all relevant critical values, and the one that resulted in the highest F-measure for the training set was chosen for the experiments.[4]

		ALL	HF	MF	LF
BMP	P	49.29	66.43	38.46	15.00
	R	47.76	59.01	39.22	17.65
	F	$\boxed{\textbf{48.52}}$	$\boxed{\textbf{62.50}}$	$\boxed{\textbf{38.83}}$	**16.22**
p_1	P	39.72	47.10	27.45	10.71
	R	50.52	75.78	41.18	17.65
	F	**44.48**	**58.10**	**32.94**	**13.33**
t	P	55.03	66.41	37.14	25.00
	R	31.96	52.80	38.24	35.29
	F	**40.43**	**58.82**	**37.68**	**29.27**
LR	P	50.71	47.11	41.67	30.00
	R	24.40	65.84	34.31	35.29
	F	**32.95**	**54.93**	**37.63**	$\boxed{\textbf{32.43}}$
$p_1 - p_2$	P	32.09	49.08	25.77	10.71
	R	47.42	66.46	41.18	17.65
	F	**38.28**	**56.46**	**31.70**	**13.33**
p_1/p_2	P	4.86	36.93	24.39	1.88
	R	70.79	70.19	39.22	41.18
	F	**9.08**	**48.39**	**30.08**	**3.60**
BASELINE	P	4.89	20.30	9.47	1.80
	R	77.32	91.93	67.65	47.06
	F	**9.20**	**33.26**	**16.61**	**3.46**

Table 1: Precision, recall, and F-measure for the four experiments. Boxes indicate the best-performing statistic in each of the cue sets.

5. Discussion

Our initial theoretical prediction, in line with (Korhonen, 2002), was that the SOA statistics should perform visibly worse than the MSC measures. The reason for this is that even though used considerably in the literature, they seem not to test for the right thing: deciding whether F is a valid frame for V should not, in principle, be based on how often it occurs with V in comparison to other verbs, as in this way, frames which are very rare for V, but otherwise common, would always be flagged as invalid. The principal source of error is the parsing procedure, when an actual frame observed in the corpus is classified as some other, possibly invalid frame. The rate of such misclassification

[3]The authors are grateful to Prof. Świdziński for making this dictionary available to them.

[4]Note that the fact that critical values were trained for each category separately is the reason why for some statistics, the overall F-measure might be larger than that for some other statistic, while the F-values for the three cue subsets are all lower. This means that the statistics 'adapts' better to the smaller categories, while giving a worse fit for the complete dataset.

should, however, be proportional to the number of classified clauses, and there seems to be no apparent relation between such error and the relation between frame frequencies in two conditional distributions. In this vein, the SOA statistics should be seen as operating on the level of some rough approximation of the actual error, most probably dependent, but definitely not directly conditioned by the true error variable.

These predictions were confirmed by BMP performing the best in three out of four categories, and p_1 performing surprisingly well in the mid- to high-frequency range despite its simplicity. The most serious discrepancy in this pattern, that is *t* and LR performing significantly better on the LF cue set, supports an unconfirmed suspicion expressed in (Korhonen et al., 2000) that these statistics should be particularly applicable to low-frequency data.

A surprising set of results is provided by the performance of the two baseline SOA statistics. p_1/p_2 performs worse than even the baseline, while $p_1 - p_2$ does astonishingly well. The probable reason for the former is that p_1/p_2 is extremely sensitive to low-frequency verbs — the less frequent frames occurring with such verbs yield very high values of p_1/p_2, thus significantly upsetting the desired ordering, where valid frames are ranked higher than invalid ones. The reason for $p_1 - p_2$ performing this well, on the other hand, is that the less frequent the frames it is applied to, the more it generally approximates p_1, and the conditional frequency comparison effect is diminished; the value of p_1 is usually much larger than that of p_2, and the difference is even more significant for rare (and in particular spurious) frames.

6. Comparison

Similar comparisons of common statistics for subcategorization acquisition can be found in the literature. (Lapata, 1999) mentions in passing that BMP, with the B_F (cf. §3.2.1.) established separately for each frame on the basis of the information contained in the COMLEX subcategorization dictionary (Grishman et al., 1994), gives results comparable to, but slightly worse than, p_1.[5] (Sarkar and Zeman, 2000; Zeman and Sarkar, 2000) compare BMP, *t* and LR and report identical results for *t* and LR, with BMP giving worse recall and better precision and F-measure, but it is not clear how reliable these results are, given various errors in the formulae for *t* and LR,[6] and given the unexplained discrepancy between the recall numbers reported in the two papers.

Finally, in a series of papers (see (Korhonen, 2002) and references therein), Korhonen carefully compares BMP, LR and p_1, using an estimate of B_F proposed by (Briscoe and Carroll, 1997), i.e., an estimate based to some extent on the unconditional probabilities of frames. This means that her version of BMP is not a pure MSC statistic in the sense of §3.2., but should rather be classified as a SOA statistic, cf. §3.3.. Korhonen notes that, of the three statistics that she compares, p_1 performs better than BMP and much better than LR, with the respective F-measures being 65.2, 53.3 and 45.1. In case of high frequency frames (above 0.01 relative frequency), p_1 results in a number of false positives similar to BMP and LR, but a much smaller number of false negatives, which implies a much higher recall. On the other hand, in case of lower frequency frames, BMP and LR show a much higher number of false positives than in the case of p_1. Similarly to the thesis of the current paper, (Korhonen et al., 2000) explain both differences by noticing that BMP and LR, but not p_1, refer not only to frame frequencies for a given verb, but also to estimates of unconditional frame probability.

[5] While BMP performs better than p_1 in the experiments reported here, with different sets of verbs for training and for the evaluation (cf. §4.), the F-measure for p_1 is higher than that for BMP if the same set is used for both tasks.

[6] In both papers, their formula for LR mentions $L(p, k_1, n_2)$ instead of $L(p, k_1, n_1)$, while their formula for *t* assumes a wrong formula for variance and lacks normalization in the denominator. Interestingly, these errors are also present in (Maragoudakis et al., 2000) and — the former — in (Sarkar and Tripasai, 2002). Moreover, the authors make no mention of distinguishing between likelihood ratios generated with positive and negative values of $p_1 - p_2$.

7. References

Brent, Michael R., 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Briscoe, Ted and John Carroll, 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*.

Dębowski, Łukasz, 2004. Trigram morphosyntactic tagger for Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski (eds.), *Intelligent Information Processing and Web Mining*, Advances in Soft Computing. Berlin: Springer-Verlag, pages 409–413.

Grishman, Ralph, Catherine Macleod, and Adam Meyers, 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of COLING '94*. Kyoto, Japan.

Korhonen, Anna, 2002. *Subcategorization Acquisition*. Ph. D. dissertation, University of Cambridge.

Korhonen, Anna, Genevieve Gorrell, and Diana McCarthy, 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of SIGDAT 2000, Hong Kong, 7–8 October, 2000*. ACL.

Lapata, Maria, 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Meeting of the North American Chapter of the Association*. College Park, MD.

Maragoudakis, Manolis, Katia Lida Kermanidis, and George Kokkinakis, 2000. Learning subcategorization frames from corpora: A case study for Modern Greek. In *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*.

Przepiórkowski, Adam, 2004. *The IPI PAN Corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski, Adam and Marcin Woliński, 2003. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings LINC-03, EACL 2003*.

Sarkar, Anoop and Woottiporn Tripasai, 2002. Learning verb argument structure from minimally annotated corpora. In *Proceedings of COLING 2002*.

Sarkar, Anoop and Daniel Zeman, 2000. Automatic extraction of subcategorization frames for Czech. In *Proceedings of COLING 2000*.

Świdziński, Marek, 1998. Syntactic dictionary of Polish verbs. Ms., Version 3a, University of Warsaw.

Zeman, Daniel and Anoop Sarkar, 2000. Learning verb subcategorization from corpora: Counting frame subsets. In *Proceedings of LREC 2000*.