Baseline Experiments in the Extraction of Polish Valence Frames

Adam Przepiórkowski[1] and Jakub Fast[2]

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

Abstract. Initial experiments in learning valence (subcategorisation) frames of Polish verbs from a morphosyntactically annotated corpus are reported here. The learning algorithm consists of a linguistic module, responsible for very simple shallow parsing of the input text (nominal and prepositional phrase recognition) and for the identification of valence frame cues (hypotheses), and a statistical module which implements three well-known inferential statistics (likelihood ratio, *t* test, binomial miscue probability test). The results of the three statistics are evaluated and compared with a baseline approach of selecting frames on the basis of the relative frequencies of frame/verb co-occurrences. The results, while clearly reflecting the many deficiencies of the linguistic analysis and the inadequacy of the statistical measures employed here for a free word order language rich in ellipsis and morphosyntactic syncretisms, are nevertheless promising.

1 Introduction

The aim of this paper is to report the results of preliminary experiments in the extraction of verbal valence frames, i.e., lists of verbs' arguments, from a corpus annotated morphosyntactically at word level.

Valence dictionaries are crucial resources for the operation of syntactic parsers, and yet, for many languages such resources are unavailable or they are available in paper form only. In case of Polish, there is only one large valence dictionary, namely, [5], although valence information is present also in the general dictionary of Polish [1]. Both dictionaries are available only in the traditional paper form. Additionally, there are two much smaller electronic valence dictionaries, created by Marek Świdziński and Zygmunt Vetulani, of about 1500 and 150 verbs, respectively.

The lack of large machine-readable valence dictionaries for Polish notwithstanding, there are many well-known arguments for constructing such dictionaries automatically, on the basis of naturally occurring texts. First of all, automatic methods of constructing valence dictionaries are much quicker and cheaper than the traditional manual process (e.g., the five volumes of [5] were published in the space of twelve years). Second, automatic methods are more objective than the traditional methods, based on potentially inconsistent intuitions of a team of lexicographers. Third, automatic methods may provide not only the categorical information, but also statistical information about how often a verb occurs with a given frame, which is particularly useful for probabilistic parsers. Fourth, the same methodology may be applied, without any overheads, to different collections of texts, e.g., to create thematic or diachronic valence dictionaries. Moreover, automatic methods may be and have been used for extending and verifying existing valence dictionaries.

The textual material for the experiments reported in this paper was the IPI PAN Corpus of Polish [7], 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), two smaller corpora were used in the experiments: the 15-million segment (over 12 million orthographic words; punctuation marks and, in some special cases, clitic-like elements are treated as separate segments) `sample` corpus, as well as a less balanced 70-million segment `wstepny` subcorpus. 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 (cf. [8]). Morphosyntactic analyses are disambiguated by a statistical tagger with a rather high 9.4% error rate [3], but all the original tags provided by the morphosyntactic analyser are also retained in the corpus.

As in the case of similar experiments reported for English, German and other languages since [2] and [4], the current algorithm consists of two stages. First, a linguistic module identifies cues, i.e., observations of co-occurrences of verbs and apparent valence frames. This stage produces much noise, partly due to the low quality of the tagger, and partly because of various deficiencies of the shallow syntactic parser used for identifying nominal phrases (NPs) and prepositional phrases (PPs). Second, the statistical module applies inferential statistics to the output of the linguistic module, trying to determine which of the valence frames observed with a given verb can be, with a certain degree of confidence, classified as actual valence frames of this verb.

The rest of the paper is structured as follows. Section 2 describes the linguistic module of the learning algorithm, i.e., shallow syntactic processing and cue identification, while the next section, 3, briefly introduces the statistics employed here. The following section, 4, describes the experiments, presents their results and evaluates them. Finally, §5 concludes the article.

2 Syntactic Cues

The process of collecting valence cues consists of three 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. Third, for each clause, all NPs and PPs identified in this clause are collected into an observed frame (OF), and the pair $\langle$verb, OF$\rangle$ is added to the set of observations. This section presents the details of the first two steps.

The shallow grammar employed in these experiments is particularly simple. Conceptually, it is a cascade of regular grammars with some added unification-like functionality for handling NP- and PP-internal agreement. There are only 9 rules for NPs and 4 rules for PPs. 5 of the NP rules are responsible for identifying very simple NPs containing a noun and possibly a sequence of pre- or post-modifying adjectives, and another 4 rules take care of personal pronouns and the strong reflexive pronoun SIEBIE. Moreover, 5 rules identify verbs, determining whether they are reflexive or negated; since in Polish valence to some extent depends on polarity, it makes sense to treat affirmative and negative verbal forms separately. Adjectival participles and gerundial forms are not considered to be verbal by this grammar.

The range of phrases that this grammar identifies is very limited: numeral phrases, adjectival phrases, adverbial phrases, clauses and infinitival verbal phrases are excluded from the consideration here, i.e., the task at hand is constrained to the identification of possible NP and PP arguments. Moreover, many NPs and PPs remain unidentified due to the fact that a very conservative approach to corpus annotation was adopted: only those forms are taken to be nominal, verbal, etc., whose all morphosyntactic interpretations are, respectively, nominal, verbal, etc. That is, the decisions made by the tagger are initially ignored. In the process of identifying NPs and PPs, all morphosyntactic interpretations of all forms belonging to an NP are unified; for example, if an adjective is syncretic between neuter singular and feminine plural, and a noun is syncretic between neuter singular and neuter plural, only the neuter singular interpretation is selected for the resulting NP. The outcome of the tagger is taken into account only in cases when this procedure results in a number of possible interpretations.

Also the second step of cue identification, i.e., finding minimal clauses, is particularly simple. Sentences are split into potential clauses on commas and conjunctions (disregarding those which have already been classified as belonging to an NP or a PP), and those potential clauses which contain a verb are selected as actual clauses. This, again, results in some noise.

Two versions of the grammar were used in the experiments, with the main difference between them being the treatment of genitive NPs (GNPs). As is well known, in Polish, GNPs often occur as modifiers of other NPs, which often results in attachment ambiguity. The shallow grammar employed here cannot resolve such ambiguities. If such GNPs were left in the output of the grammar, this would result in many false observed valencies involving GNPs. Both grammars currently deal with this problem via brute force: the first grammar (G1) ignores GNPs altogether (so there is no way to identify valence frames with genuine GNP arguments), while the second grammar (G2) attaches GNPs to immediately proceeding NPs and PPs, whenever possible.

The final simplification assumed here is that nominative NPs are ignored altogether, i.e., no attempt at distinguishing subject-taking verbs and subjectless verbs is made.

3 Basic Statistics

Three standard statistics were compared for rating each observed verb/frame combination in terms of how strong a piece of evidence it is for inferring the verb's valence requirements. The statistics were: binomial miscue probability testing, the *t* test for non-normal variables, and Likelihood Ratio.

The three metrics employ a common probabilistic model. The values characterising a given verb/frame combination $\langle v, f \rangle$ are interpreted as describing two binomial samples, m_X and m_Y, such that in m_X, the number of successes k_X is equal to the number of f's occurrences with v and the sample size n_X is equal to the total number of v's occurrences, and in m_Y, the number of succeseses k_Y is the number of occurences of f in clauses that *do not* contain v, and sample size n_Y is the total number of such clauses. The samples are interpreted as taken from two binomially-distributed random variables $X \sim \mathrm{Bin}(n_X, \pi_X)$ and $Y \sim \mathrm{Bin}(n_Y, \pi_Y)$, where π_X and π_Y are estimated on the basis of m_X and m_Y as $\hat{\pi}_X = \frac{k_X}{n_X}$ and $\hat{\pi}_Y = \frac{k_Y}{n_Y}$.

3.1 Binomial Miscue Probability Test

Assuming a certain maximal probability B_f that — due to parser or grammatical error — a given frame f occurs in a sentence irrespective of the verb it contains (in the experiments, B_f was set to 2^{-7}), the binomial test measures the probability of k_X or more occurences of f being generated in a sample of size n_X by a random variable with a distribution $\mathrm{Bin}(n_X, B_s)$. Its value is calculated as follows:

$$H(B_f) = \sum_{i=k_X}^{n_X} (B_f)^i \, (1 - B_f)^{n_X - i} \binom{n_X}{i} \tag{1}$$

If this probability is sufficiently low (i.e., lower than an assumed null hypothesis rejection level), the probability that the observed number of co-occurrences is due solely to error can be neglected and the frame classified as valid for the verb.

3.2 *t* Test

The *t* test establishes the significance of a difference observed between the probabilities of success in two independent samples. For binomially-distributed variables, its value is given by the following equation:

$$t = \frac{\hat{\pi}_X - \hat{\pi}_Y}{\sqrt{\frac{\hat{\sigma}^2_X}{n_X} + \frac{\hat{\sigma}^2_Y}{n_Y}}} \tag{2}$$

where $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$ and is the variance in a single Bernoulli trial estimated on the basis of the respective samples. The null hypothesis for this test is that the two samples come from the same population, and the alternative hypothesis is that m_X comes from a population characterised by a significantly lower probability of success. If the value of t is lower than an appropriate threshold value, it can be inferred that the probability of succes in m_X is significantly lower than that in m_Y, the null hypothesis is rejected and f is taken to be an invalid frame for v.

3.3 Likelihood Ratio

The Likelihood Ratio test measures the significance of the difference in the probability of a particular outcome given two different, nested classes of probabilistic models. Assuming a joint distribution $\langle X, Y\rangle$, the likelihood ratio is used to compare the quality of the best model in which X and Y have the same probabilities of success against the quality of the best joint binomial model that does not make this assumption. The value of the likelihood ratio is thus the following:

$$\lambda = \frac{\max_p P(\langle m_X, m_Y\rangle | \langle X, Y\rangle \sim \mathrm{Bin}(n_X, p) \times \mathrm{Bin}(n_Y, p))}{\max_{p_X, p_Y} P\left(\langle m_X, m_Y\rangle | \langle X, Y\rangle \sim \mathrm{Bin}(n_X, p_X) \times \mathrm{Bin}(n_Y, p_Y)\right)} \tag{3}$$

A sufficiently low value of λ implies that the theoretical values of π_X and π_Y are sufficiently different. In the experiments described below, λ was multiplied by -1 if the difference $p_X - p_Y$ was negative, in order to eliminate only such $\langle v, f\rangle$ combinations where f is significantly less likely to occur with v than with any other verb, and not the other way round. If the value of λ is negative and sufficiently close to 0, it can be concluded that f is an invalid frame for v.

4 Experiments and Evaluation

Altogether 16 experiments were performed: for the two subcorpora mentioned in §1, `sample` and `wstepny`, two versions of the grammar described in §2, G1 and G2, were used, and for each of the corpus/grammar combinations, four statistics were applied: the three statistics described in §3, as well as the baseline statistic, i.e., the Maximum Likelihood Estimate (MLE) of π_X, where frames with $\hat{\pi}_X \geq 0.01$ were selected as valid. The result of each of the experiments is a function from verb lemmata to sets of valence frames accepted for those verbs by a given statistic, on the basis of cues generated by a given grammar running on a given corpus. Frames are understood as multisets of arguments.

For example, the following 8 frames were selected for the verb DOSTRZEC (*discern*) by the binomial miscue probability statistic, on the basis of data generated by G2 from the `wstepny` corpus: `<empty>` (empty valence frame), `np/acc` (an accusative NP), `pp/w/loc` (a PP consisting of the preposition W and a locative NP), `np/acc+np/gen`, `np/acc+pp/w/loc`, `np/acc+pp/w/loc`, `np/gen`, `np/gen+pp/w/loc`. Valence dictionaries used for the evaluation (see below) mention only one valence frame, `np/acc`, but — with the possible exception of `np/acc+np/gen` — other discovered valencies are reasonable sets of dependents of DOSTRZEC, with `pp/w/loc` being a common locative modifier of DOSTRZEC and `np/gen` being a possible effect of the (long-distance) genitive of negation or genitive modifiers, and with the observations `<empty>` and `pp/w/loc` reflecting the ellipsis of the object. 38 other observed frames for DOSTRZEC were rejected by this test.

For the purpose of the experiments reported here, the confidence level for all statistics was set to 99%. Moreover, only frames registered a certain minimal number of times were taken into account, and additionally a certain minimum verb/frame co-occurrence restriction was imposed. Four such ⟨frame occurrences, frame/verb co-occurrences⟩ cutoff points were tested, from $\langle 50, 3\rangle$ to $\langle 800, 10\rangle$.

Two 'gold standards' were adopted for the purpose of evaluating the results of these experiments, namely, the two dictionaries containing valence information mentioned in §1: [5] and [1]. From the set of verbs for which at least 100 observations were registered by the G1 grammar in the `sample` corpus, 48 verbs[1] were blindly selected, evenly distributed across the scale of the number of occurrences. For each of these verbs, valence information was manually extracted from the two valence dictionaries. It should be noted that in both cases, but perhaps especially in the case of [5], this was an interpretative process, due to the fact that both dictionaries refer to some of the arguments via their function (e.g., a temporal phrase) and not their morphosyntactic form. Those valence frames were narrowed down to valence frames containing only phrases identifiable by the grammars, i.e., NPs and PPs.

Bańko [1]

	`sample`		`wstepny`	
	G1	G2	G1	G2
MLE	40.17	42.23	38.33	36.94
binomial	**40.96**	42.80	**38.94**	**38.01**
t test	39.83	43.90	36.00	35.24
likelihood	39.93	**44.23**	35.69	35.06

Polański [5]

	`sample`		`wstepny`	
	G1	G2	G1	G2
MLE	40.07	42.32	37.35	36.47
binomial	**41.13**	43.28	**38.42**	**37.86**
t test	39.08	42.92	33.23	32.54
likelihood	39.11	**43.41**	33.48	32.65

Table 1. F values for the 16 tests, for two gold standards (based on [1] and [5]), for the cutoff point $\langle 400, 10\rangle$.

[1] BIĆ SIĘ, DOBIEC, DOCZEKAĆ SIĘ, DOSTRZEC, DŹWIGAĆ, NABIERAĆ, NAUCZYĆ SIĘ, OBSŁUGIWAĆ, ODBIJAĆ SIĘ, ODCZYTYWAĆ, OKAZAĆ SIĘ, OPOWIADAĆ, ORZEKAĆ, PŁAKAĆ, POPATRZEĆ, POSTĘPOWAĆ, POTRZEBOWAĆ, POWIĘKSZYĆ, PRZESTRZEGAĆ, PRZEŻYWAĆ, REALIZOWAĆ, ROZEGRAĆ, SIĄŚĆ, SPACEROWAĆ, STARTOWAĆ, SZANOWAĆ, UCZESTNICZYĆ, UDAWAĆ SIĘ, URZĄDZAĆ, USTANOWIĆ, WKRACZAĆ, WYBIERAĆ SIĘ, WYNIKAĆ, WYSTAWIAĆ, WYTWARZAĆ, WZDYCHAĆ, ZABRZMIEĆ, ZAJMOWAĆ, ZAJRZEĆ, ZAPOWIADAĆ, ZATRZYMYWAĆ SIĘ, ZAWIERAĆ, ZAŻĄDAĆ, ZDARZAĆ SIĘ, ZDAWAĆ SIĘ, ZJECHAĆ, ZRZUCIĆ, ZWRACAĆ.

Table 1 presents the summary of the results for the two gold standards. The numbers in the table are the average values of F-measure (the harmonic mean of precision and recall) for the 48 verbs, where only those frames were taken into account which were observed at least 400 times, and co-occurred with a given verb at least 10 times. The best statistic for each corpus/grammar pair is indicated by bold face.

As can be seen in Table 1, the binomial miscue probability test is the best statistic in six of the eight corpus/grammar/gold standard combinations. As can be seen from the comparison of Tables 1 and 2, this test also turns out to be the most stable, i.e., least dependent on cutoff points. Another interesting observation is that, surprisingly, the results for the much smaller `sample` corpus are uniformly better than the results for `wstepny`. The decrease in F-measures is almost entirely due to reduced precision; for example, the 41.89/35.99 difference in Table 2 for the binomial test applied to the results of G2 as evaluated on the basis of [1] corresponds to recall values of 74.07/73.72 and precision values of 34.31/27.37. Why this should be so is still not fully clear. Finally, it should be noted that the baseline MLE test mentioned above almost always fares better than the more sophisticated *t* test and likelihood ratio.

Bańko [1]

	`sample`		`wstepny`	
	G1	G2	G1	G2
MLE	34.77	35.47	31.16	29.87
binomial	**41.79**	**41.89**	**38.14**	**35.99**
t test	30.28	32.32	22.31	18.62
likelihood	30.28	32.69	22.01	18.57

Polański [5]

	`sample`		`wstepny`	
	G1	G2	G1	G2
MLE	36.02	36.37	32.55	31.39
binomial	**40.21**	**41.67**	**37.52**	**37.83**
t test	30.50	32.15	23.06	20.65
likelihood	30.92	32.51	23.00	20.51

Table 2. As Table 1, but for the cutoff point $\langle 50, 3 \rangle$.

The numbers in Table 1 may appear to be disappointing, especially as compared with F-measures given in the literature, e.g., F-measures of around 80 reported in [9]. However, these numbers are not comparable, as the results usually reported in the literature are F-measures based on so-called token precision and recall, i.e., the evaluation is based on the manual annotation of frames in a test corpus, usually performed by the authors. In the current paper, on the other hand, automatically extracted frames are evaluated on the basis of two valence dictionaries given *a priori*. Moreover, the two dictionaries differ significantly in the kind of valence information they contain. In fact, when the information in one dictionary is evaluated against the other, the F-measure is around 65.5, so — in a sense — this is the upper limit of what can be achieved using the methodology adopted here.[2]

[2] For [5] treated as extracted data and [1] as gold standard, the average precision, recall and F for the 48 verbs (narrowed down to frames consisting of NPs and PPs

5 Conclusion

To the best of our knowledge, the experiments reported here are the first attempt at the extraction of valence frames for a free word order Slavic language from a corpus containing no syntactic constituency annotation, rather than from a treebank, cf., e.g., [9].

The manual inspection of the results suggests that many 'errors' stem from the phenomenon of ellipsis, from erroneous classification of adjuncts as arguments, and from various errors and inconsistencies in gold standards. Taking into consideration the fact that the cross-gold standard agreement, as given by the F-measure, is around 65.5, and given the high noise production at each level of linguistic processing (morphological analysis, tagging, shallow parsing) and the basic nature of statistical models involved, the results reported here are highly encouraging.

References

1. Mirosław Bańko, editor. *Inny słownik języka polskiego.* Wydawnictwo Naukowe PWN, Warsaw, 2000.
2. Michael R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262, 1993.
3. Łukasz Dębowski. Trigram morphosyntactic tagger for Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 409–413. Springer-Verlag, Berlin, 2004.
4. Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st ACL*, pages 235–242, 1993.
5. Kazimierz Polański, editor. *Słownik syntaktyczno-generatywny czasowników polskich.* Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Kraków, 1980–1992.
6. Adam Przepiórkowski. On the computational usability of valence dictionaries for Polish. IPI PAN Research Report 971, Institute of Computer Science, Polish Academy of Sciences, 2003.
7. Adam Przepiórkowski. *The IPI PAN Corpus: Preliminary version.* Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2004.
8. Adam Przepiórkowski and Marcin Woliński. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*, pages 109–116, 2003.
9. Anoop Sarkar and Daniel Zeman. Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 691–697, 2000.

only) are 74.77, 65.78 and 64.54, respectively, and for the opposite evaluation — 65.65, 76.59 and 64.36.