New methods for verb valence extraction from Polish texts

Łukasz Dębowski, Marcin Woliński
{ldebowsk,wolinski}@ipipan.waw.pl

Institute of Computer Science, Polish Academy of Sciences

The problem

- Various resources are needed to parse a text in a natural language: grammars, dictionaries, etc.

> Jasio podziękował Marysi za współpracę.
> *(Johnny thanked Mary for cooperation.)*
> Marysia odpowiedziała mu, że nie ma za co.
> *(Mary answered him: "Not at all".)*

- The valence dictionary is one of them.

> *odpowiedzieć*: np(nom), np(dat), ZE
> *podziękować*: np(nom), np(dat), za+np(acc)

- Several valence dictionaries for Polish have been published.
- Is it possible to extract a better dictionary from a corpus?

Valence frames

Johny was somewhat surprised by the answer...

Odpowiedź dziwiła Jasia dwa tygodnie.
Jasia dziwiło, że Marysia tak odpowiedziała.
Jasio dziwił się i dziwił.
Jasio dziwił się i Marysi, i odpowiedzi.
Jasio dziwił się także, że się tak długo dziwi.

$$\mathbf{F}(\textit{dziwić}) = \left\{ \begin{array}{l} \{\mathrm{np(nom)}, \mathrm{np(acc)}\}, \\ \{\mathrm{ZE}, \mathrm{np(acc)}\}, \\ \{\mathrm{np(nom)}, \mathrm{sie}\}, \\ \{\mathrm{np(nom)}, \mathrm{sie}, \mathrm{np(dat)}\}, \\ \{\mathrm{np(nom)}, \mathrm{sie}, \mathrm{ZE}\} \end{array} \right\}$$

Valence is hard to describe

1. Certain arguments must not co-occur.
2. It is difficult to enumerate all frames.
3. Some arguments can be dropped.
4. Some arguments are required.
5. The argument/adjunct distinction is vague:

> Odpowiedź brzmiała dziwacznie.

6. Equivalence classes of arguments depend on the verb:

> Jasio dziwił się { że Marysia tak odpowiedziała. / jej odpowiedzi.

Brent (1993)

Input: a corpus of raw texts,
a deterministic shallow parser.

$\mathbf{c(v, f)}$ — the count of parses that contain frame $\mathbf{f}$ and verb $\mathbf{v}$,
$\mathbf{c(v)} = \sum_{\mathbf{f}} \mathbf{c(v, f)}$ — the count of parses that contain $\mathbf{v}$.

We will consider $\mathbf{f}$ a frame for $\mathbf{v}$ if and only if

$$\sum_{\mathbf{n=c(v,f)}}^{\mathbf{c(v)}} \binom{\mathbf{c(v)}}{\mathbf{n}} \mathbf{p_f^n (1 - p_f)^{c(v)-n}} \leq \mathbf{0.05},$$

where $\mathbf{p_f}$ is selected in supervised or unsupervised learning:

1. to minimze the error rate on a training dictionary,
2. so that binomial distribution $\mathbf{B(\cdot, p_f)}$ fits into the first cluster in the histogram of verbs grouped by rate $\mathbf{c(v, f)/c(v)}$.

The scheme of our approach

1. The Świgra parser is used to analyse raw text.
2. Parse forests are reduced to sets of valence frames.
3. The sets are disambiguated with a new EM selection algorithm.
4. The occurrences of frames and verbs are counted.
5. Supervised learning is used to find the sets of possible and required arguments.
6. The sets of valence frames are corrected in the next step.

Parsing

The sentences were selected from the IPI PAN Corpus of Polish.

We used a permissive version of the Świgra parser:

- $\leq$ **1** subject and any number of other arguments.

We selected sentences that

- contained $\leq$ **15** words,
- could be analysed in $\leq$ **1** minute,
- had $\leq$ **40** parses per elementary sentence.

Only $\leq$ **5000** sentences per verb were considered.

The reduction of parse trees to valence frames

We deleted parses that contained:

- *to*, *co*, *nic* (=*this*, *what*, *nothing*),
- highly improbable interpretations of ambiguous words.

The remaining parses were transformed:

- Phrases that are not children of the verb were deleted.
- The dropped personal subject was added.
- The genitive of negation was treated.
- Some NPs and PPs were marked as adjuncts.
- Lemmas of the phrase heads were removed.

The bank of reduced parse forests

- The bank consists of **510 743** elementary sentences.
- Świgra has processed about **2.6** times more sentences.
- About **3.3** million running words.

```
'Kto zastąpi piekarza?' ('Who will replace the baker?')
zastąpić :np:acc: :np:nom:
zastąpić :np:gen: :np:nom:
'Nie płakał na podium.' ('He did not cry on the podium')
płakać :np:nom: :prepnp:na:acc:
płakać :np:nom: :prepnp:na:loc:
```

The EM selection

$\mathbf{Y_i}$ — the set of parses for the $\mathbf{i}$-th sentence, $\mathbf{i = 1, 2, ..., M}$.
$\mathbf{p_j^{(n)}}$ — the frequency of parse $\mathbf{j}$ in the $\mathbf{n}$-th iteration, $\mathbf{p_j^{(1)} = 1}$.

$$\mathbf{p_{ji}^{(n)}} = \begin{cases} \mathbf{p_j^{(n)}} / \sum_{\mathbf{j' \in Y_i}} \mathbf{p_{j'}^{(n)}}, & \mathbf{j \in Y_i}, \\ \mathbf{0}, & \text{else}, \end{cases}$$

$$\mathbf{p_j^{(n+1)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{p_{ji}^{(n)}}.$$

sampling	accuracy
the shortest frame with maximal $\mathbf{p_{ji}^{(n)}}$	**72.6%**
the frame with maximal $\mathbf{p_{ji}^{(n)}}$	**72.4%**
the shortest frame	**57.5%**
blind	**46.9%**

Tested on 190 sentences, 500 Monte Carlo simulations, $\mathbf{n = 10}$.

Other applications of the EM selection

The EM selection may be used in many NLP applications.
E.g., for morphological disambiguation without a training set:

$\mathbf{Y_i}$ — possible interpretations for the $\mathbf{i}$-th token, $\mathbf{i = 1, 2, ..., M}$.
$\mathbf{p_j^{(n)}}$ — the frequency of interpretation $\mathbf{j}$ in the $\mathbf{n}$-th iteration.

$$\mathbf{p_{ji}^{(n)}} = \begin{cases} \mathbf{p_j^{(n)}} / \sum_{\mathbf{j' \in Y_i}} \mathbf{p_{j'}^{(n)}}, & \mathbf{j \in Y_i}, \\ \mathbf{0}, & \text{else}, \end{cases}$$

$$\mathbf{p_j^{(n+1)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{p_{ji}^{(n)}}.$$

$$\mathbf{L^{(n+1)} \geq L^{(n)}} \text{ for } \mathbf{L^{(n)}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} \mathbf{\log} \left[\sum_{\mathbf{j \in Y_i}} \mathbf{p_j^{(n)}} \right].$$

The local maxima form a convex set so setting $\mathbf{p_j^{(1)}}$ doesn't matter.

The proto-dictionary, the training data, and the test data

The proto-dictionary (counts of parses in the disambiguated bank):

```
'przyłapać' => { ('to catch smb red-handed')
   'np(acc),np(gen),np(nom)' => 1,
   'na+np(loc),np(nom),sie' => 1,
   'na+np(loc),np(gen),np(nom)' => 1,
   'np(acc),np(nom)' => 4,
   'adv,np(nom)' => 1,
   'na+np(loc),np(acc),np(nom)' => 3
}
```

The training set:

- valence frames for **1166** verbs according to Świdziński (1994).

The test set:

- valence frames for **203** verbs according to the dictionaries by Polański (1980), Świdziński (1994), and Bańko (2000).

A new description of verb valence

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

We approximate **F(v)** with three objects:

1. **L(v)** — the set of possible arguments for **v**,
2. **E(v)** — the set of required arguments for **v**,
3. $\mathbf{M(v)} : \mathbf{L(v)} \times \mathbf{L(v)} \longrightarrow \{\leftarrow, \rightarrow, \leftrightarrow, \times, \perp\}$
   — co-occurrence matrix for the arguments of **v**.

The cells of the co-occurrence matrix depend weakly on the verb.

The sets of arguments

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

$$\mathbf{L(v)} := \bigcup_{\mathbf{f} \in \mathbf{F(v)}} \mathbf{f}$$

$$\mathbf{E(v)} := \bigcap_{\mathbf{f} \in \mathbf{F(v)}} \mathbf{f}$$

The co-occurence matrix

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

Let $\langle \mathbf{a} \rangle := \{\mathbf{f} \in \mathbf{F(v)} \mid \mathbf{a} \in \mathbf{f}\}$ and

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &\iff \langle \mathbf{a} \rangle \cap \langle \mathbf{b} \rangle = \emptyset, && (\mathbf{a} \text{ excludes } \mathbf{b}) \\ \mathbf{a} \leftrightarrow \mathbf{b} &\iff \langle \mathbf{a} \rangle = \langle \mathbf{b} \rangle, && (\mathbf{a} \text{ and } \mathbf{b} \text{ co-occur}) \\ \mathbf{a} \rightarrow \mathbf{b} &\iff [\langle \mathbf{a} \rangle \subset \langle \mathbf{b} \rangle \wedge \langle \mathbf{a} \rangle \neq \langle \mathbf{b} \rangle], && (\mathbf{a} \text{ implies } \mathbf{b}) \\ \mathbf{a} \leftarrow \mathbf{b} &\iff [\langle \mathbf{a} \rangle \supset \langle \mathbf{b} \rangle \wedge \langle \mathbf{a} \rangle \neq \langle \mathbf{b} \rangle], && (\mathbf{b} \text{ implies } \mathbf{a}) \\ \mathbf{a} \perp \mathbf{b} &\iff \langle \mathbf{a} \rangle \setminus \langle \mathbf{b} \rangle, \langle \mathbf{a} \rangle \cap \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle \setminus \langle \mathbf{a} \rangle \neq \emptyset. \\ & && (\mathbf{a} \text{ and } \mathbf{b} \text{ are formally independent}) \end{aligned}$$

$$\mathbf{M(v)_{ab}} := \mathrm{R} \iff \mathbf{a}\,\mathrm{R}\,\mathbf{b}$$

An example

$$\mathbf{F}(\textit{przyłapać}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{np(acc)}, \text{na+np(loc)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{na+np(loc)}\} \end{array} \right\}$$

M(*przyłapać*)	np(nom)	np(acc)	sie	na+np(loc)
np(nom)	$\leftarrow\!\!\!\!\uparrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
np(acc)	$\uparrow$	$\leftarrow\!\!\!\!\uparrow$	$\times$	$\bot$
sie	$\uparrow$	$\times$	$\leftarrow\!\!\!\!\uparrow$	$\bot$
na+np(loc)	$\uparrow$	$\bot$	$\bot$	$\leftarrow\!\!\!\!\uparrow$

A comparison of matrices (Bańko vs. Świdziński)

The numbers of triplets $\mathbf{(v, a, b)}$ with the given values of $\mathbf{M(v)_{ab}}$:

		Bań.					
		×	←	→	↔	⊥	N/A
Świ.	×	364	2	2	–	30	724
	←	4	253	1	2	18	176
	→	4	1	253	2	18	176
	↔	–	25	25	410	2	124
	⊥	16	6	6	–	28	38
	N/A	1242	230	230	167	102	

High agreement rate

The numbers of triples **(v, a, b)** in the pairs of dictionaries:

	Σ	same **M(v)$_{ab}$**	different **M(v)$_{ab}$**	agreement rate
in Bań. & Pol.	1383	1187	196	86%
in Bań. & Świ.	1472	1308	164	89%
in Pol. & Świ.	1449	1283	166	89%

Valence dictionaries report quite different sets of arguments for a given verb. Nevertheless, combinations of the arguments that appear in several dictionaries are constrained similarly.

The most frequent values of matrix cells

	np(nom)	np(acc)	advp	np(dat)	np(inst)	w+np(loc)	do+np(gen)	na+np(acc)	z+np(gen)
np(acc)	↑:66%								
adv	↑:94%	↑:57%							
np(dat)	↑:91%	↑:50%	×:72%						
np(inst)	↑:98%	↑:57%	×:85%	×:64%					
w+np(loc)	↑:91%	↑:61%	×:92%	×:71%	×:85%				
do+np(gen)	↑:94%	↑:48%	×:92%	×:80%	×:89%	×:100%			
na+np(acc)	↑:96%	↑:64%	×:96%	×:85%	×:86%	×:90%	×:100%		
z+np(gen)	↑:100%	↑:65%	×:94%	×:91%	×:100%	×:100%	×:92%	×:100%	
ZE	↑:91%	×:76%	×:86%	⊥:64%	×:100%	×:100%	×:80%	×:71%	⊥:100%

$\mathbf{M(v)_{ab}}$ for fixed arguments **a** and **b** depends weakly on verb **v**.

Pruning of the proto-dictionary

1. Compute $\mathbf{L(v)}$ and $\mathbf{E(v)}$ for $\mathbf{F(v)}$ given by the proto-dict.
2. Correct $\mathbf{L(v)}$ and $\mathbf{E(v)}$. Then reconstruct

$$\mathbf{F(v)} := \{(\mathbf{f} \cup \mathbf{E(v)}) \cap \mathbf{L(v)} \mid \mathbf{f} \in \mathbf{F(v)}\}\,.$$

3. Compute $\mathbf{M(v)}$ for the current $\mathbf{F(v)}$.
4. Correct $\mathbf{M(v)}$. Then reconstruct

$$\mathbf{F(v)} := \left\{ \mathbf{f} \in 2^{\mathbf{L(v)}} \;\middle|\; \begin{array}{l} \forall_{\mathbf{a} \in \mathbf{E(v)}}\, \mathbf{a} \in \mathbf{f}, \\ \forall_{\mathbf{a},\mathbf{b} \in \mathbf{L(v)}}\, \phi(\mathbf{f}, \mathbf{M(v)}, \mathbf{a}, \mathbf{b}) \end{array} \right\},$$

where

$$\phi(\mathbf{f}, \mu, \mathbf{a}, \mathbf{b}) := \begin{cases} \neg(\mathbf{a} \in \mathbf{f} \wedge \mathbf{b} \in \mathbf{f}), & \mu_{\mathbf{ab}} = \times, \\ \mathbf{a} \in \mathbf{f} \iff \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \leftrightarrow, \\ \mathbf{a} \in \mathbf{f} \implies \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \rightarrow, \\ \mathbf{a} \in \mathbf{f} \impliedby \mathbf{b} \in \mathbf{f}, & \mu_{\mathbf{ab}} = \leftarrow, \\ \text{true}, & \mu_{\mathbf{ab}} = \perp\,. \end{cases}$$

Pruning of the argument sets

Argument **a** is recognized as possible for verb **v** if

$$\mathbf{c(v, a)} \geq \mathbf{p_a c(v)} + \mathbf{1},$$

and then **a** is recognized as required for **v** unless

$$\mathbf{c(v)} - \mathbf{c(v, a)} \geq \mathbf{p_{\neg a} c(v)} + \mathbf{1},$$

where

- $\mathbf{c(v, a)}$ — the count of parses that contain both **a** and **v**,
- $\mathbf{c(v)}$ — the count of parses that contain **v**,
- $\mathbf{p_a}$ and $\mathbf{p_{\neg a}}$ — parameters tuned to minimize the error rate on the training set.

Pruning of the co-occurrence matrices

- S — relation $\mathbf{M(v)_{ab}}$ in the raw matrix
- R — the most frequent relation between **a** and **b** in the training set across different verbs

	agreement rate
$\mathbf{M(v)_{ab}} := \mathsf{S}$	77%
$\mathbf{M(v)_{ab}} := \mathsf{R}$	80%
$\mathbf{M(v)_{ab}} := \mathsf{T}$	83%

- $\mathsf{T} = \begin{cases} \mathsf{R}, & \mathbf{C(a\,R\,b)} \geq \mathbf{p}_{\mathsf{S}\Rightarrow\mathsf{R}}\mathbf{C(ab)} + \mathbf{t}_{\mathsf{S}\Rightarrow\mathsf{R}}, \\ \mathsf{S}, & \text{otherwise}, \end{cases}$
- $\mathbf{C(ab)}$ — the number of verbs which select for **a** and **b**
- $\mathbf{C(a\,R\,b)}$ — the number of verbs for which $\mathbf{a}\,\mathsf{R}\,\mathbf{b}$
- $\mathbf{p}_{\mathsf{S}\Rightarrow\mathsf{R}}$ and $\mathbf{t}_{\mathsf{S}\Rightarrow\mathsf{R}}$ — optimized parameters

A comparison with an earlier experiment

Fast and Przepiórkowski (2005):

- valence extracted from the IPI PAN Corpus (12 mln words),
- shallow parser,
- Brent's (1993) method (BMP),
- training and testing on Świdziński's dictionary,
- only NP and PP arguments.

	REC	PRE
Fast i Przepiórkowski (2005)	**48%**	**49%**
our results (projected onto NPs and PPs)		
— proto-dictionary	**73%**	**19%**
— after argument pruning	**48%**	**64%**
— after the co-occurence matrix correction	**43%**	**74%**
— BMP reapplied	**33%**	**85%**

Some drawbacks of our method

- The distribution of false positives:

FP	all	argument climbing	test set errors
np(acc)	9	4	1
sie	4	0	4
np(dat)	11	3	6

- Our correction makes many cells of co-occurence matrices independent of the verb.
- Exclusion ZE $\times$ np(nom) that holds under the absence of *się* is atypical and impossible to describe:

$$\mathbf{F}(\textit{dziwić}) = \left\{ \begin{array}{l} \{\text{np(nom)}, \text{np(acc)}\}, \\ \{\text{ZE}, \text{np(acc)}\}, \\ \{\text{np(nom)}, \text{sie}\}, \\ \{\text{np(nom)}, \text{sie}, \text{np(dat)}\}, \\ \{\text{np(nom)}, \text{sie}, \text{ZE}\} \end{array} \right\}$$

A comparison with three test dictionaries

The numbers of pairs **(v, f)**:

The numbers of pairs **(v, a)**:

a) after the correction of co-occurence matrix:

F	AE	Bań.	Pol.	Świ.	MV
AE	658				
Bań.	418	1646			
Pol.	359	771	1519		
Świ.	363	758	770	1359	
MV	394	984	996	983	1209
recall	0.33	0.81	0.82	0.81	
precision	0.6	0.6	0.66	0.72	

L	AE	Bań.	Pol.	Świ.	MV
AE	674				
Bań.	603	1330			
Pol.	586	956	1320		
Świ.	581	899	954	1251	
MV	600	1056	1111	1054	1211
recall	0.5	0.87	0.92	0.87	
precision	0.89	0.79	0.84	0.84	

b) after reapplying Brent's supervised procedure:

F	AE	Bań.	Pol.	Świ.	MV
AE	413				
Bań.	311	1622			
Pol.	275	759	1498		
Świ.	294	743	750	1323	
MV	311	965	972	956	1178
recall	0.26	0.82	0.83	0.81	
precision	0.75	0.59	0.65	0.72	

L	AE	Bań.	Pol.	Świ.	MV
AE	582				
Bań.	524	1308			
Pol.	520	941	1296		
Świ.	521	883	930	1220	
MV	530	1038	1085	1027	1182
recall	0.45	0.88	0.92	0.87	
precision	0.91	0.79	0.84	0.84	

Thank you!