Semantic relation recognition within Polish noun phrase: A rule-based approach

Paweł Kędzia
Institute of Informatics
Wrocław University of Technology
pawel.kedzia@pwr.wroc.pl

Marek Maziarz
Institute of Informatics
Wrocław University of Technology
marek.maziarz@pwr.wroc.pl

Abstract

The paper[1] presents a rule-based approach to semantic relation recognition within the Polish noun phrase. A set of semantic relations, including some thematic relations, has been determined for the need of experiments. The method consists in two steps: first the system recognizes word pairs and triples, and then it classifies the relations. Evaluation was performed on random samples from two balanced Polish corpora.

1 Introduction

Semantic relation recognition is a well-known task in natural language processing. Although the relation recognition within noun phrase and between nominals was studied intensely, the task is still challenge for semantic analysis of Polish. We are aware of few papers and projects dealing with Semantic Role Labelling between predicates and their arguments, cf. (Gołuchowski and Przepiórkowski, 2012) or (Lun, 2009), but of none concerning semantic relation recognition inside Polish noun phrase.

2 Related work

In (Nastase et al., 2006) authors classify semantic relations between a head and a modifier of a noun phrase. Number of all relation types was equal to 30. These relations were grouped into 5 more general groups. The authors experimented with decision trees, instance-based learning and Support Vector Machines. For each relation they learnt the binary classifier; as the baseline for F-measure they used the model with all of examples classified as positive and recall being equal to 100%. With regard to the semantic relation the baseline ranged between 17.78% and 60.35%.

Identifying the semantic relations inside compound nouns was presented in (Uchiyama et al., 2008). The authors used SVM classifier and in the best configuration of features, they achieved accuracy of about 84%.

In (Rosario and Hearst, 2001) authors used neural networks to determine 20 semantic relations – similarily to (Nastase et al., 2006) – between a head and a modifier of noun phrase. They used a domain-specific lexical hierarchy of medicine. The authors achieved accuracy of about 60%.

The workshop SemEval-2010 (task 8) concerned the recognition of semantic relations between nominals. In (Tratz and Hovy, 2010) the authors developed a system based on the Maximum Entropy classifier, able to detect 10 bidirectional semantic relations Achieved F-measures depended on the system configuration and lay between $66,68\%$ and $77,75\%$. The same set of semantic relations was used in (Rink and Harabagiu, 2010). The authors used Support Vector Machines classifier and a very rich set of features (i.e., part of speech for all constituents of a semantic relation pair, number of words between the nominals, features based on paths in the dependency tree from Stanford dependency parser). F-measure of this approach was 82.19%.

Authors in (Tymoshenko and Giuliano, 2010) used shallow syntactic parsing and semantic information from ResearchCyc (Lenat, 1995) in the same task of recognizing semantic relations. They used liner combination of kernels (semantic and syntactic) using Support Vector Machines classifier. For the best combination of kernels, they obtained F-measure equal to 77.62%.

There are some works, where rule-based approaches were used. In (Huang, 2009) there has been proposed an approach for automatic construction of rules identifying ten types of seman-

[1]Work financed by The National Centre for Research and Development project SP/I/1/77065/10.

Proceedings of Recent Advances in Natural Language Processing, pages 342–349,
Hissar, Bulgaria, 7-13 September 2013.

tic relations, using five types of input informations. The relation instances were extracted from Modern Chinese Standard Dictionary. The authors achieved very high precision (range from $0,81$ to 0.99), but recall was low - about $0,2$. In (Hearst, 1992) authors used set of manually written rules for identification of hyperonymy relations. (Ben Abacha and Zweigenbaum, 2011) used linguistic patterns (built semi-automatically from corpora) to identify semantic relatios in medical texts. In this domain-specific task they achieved 75.72% precision and $60,46\%$ recall.

3 Recognized semantic relation types

We seek for semantic relations within nominal phrases. The relation set consists of 12 semantic relations, of which 5 are thematic (semantic) roles[2]. Definitions of our semantic relations are based on works of (Kearns, 2011), (Palmer et al., 2010), (Van Valin, 2004), (Larson, 1996), (Dowty, 1991), (Jędrzejko, 1993), (Laskowski and Wróbel, 1997). We tried to select relations that are very frequent or frequent in Polish texts.[3] The relation set is following (thematic roles are marked with *theta*, other relations – with *rho*):

Proto-Agent$_\theta$ – it is an instigator of an action or an entity that is in a particular state, it may undergoe change of state not caused by another participant; for predicates denoting relations – it is the first element of the relation: (człowiek) wykształcony przez Jana$_\theta$ '(man) educated by John$_\theta$', wyjący wilk$_\theta$ 'howling wolf$_\theta$'. The Proto-Agent macrorole covers subroles of Agent, Causer and non-agentive non-causative Actor (cf. *Actor* macrorole in (Kearns, 2011)).

Proto-Patient$_\theta$ is the second macrorole – it is an entity undergoing action, event or change of state caused by another participant; for predicates denoting relations – it is the second element of a given relation: wykształcenie kogoś$_\theta$ 'educating someone$_\theta$', (Jan) posiadający majątek$_\theta$ '(John) possessing an estate$_\theta$'. According to (Dowty, 1991) many thematic roles come down to the macroroles of Proto-Patient and Proto-Agent.

Instrument$_\theta$ is a tool, a device or means used by someone in order to cause something, it is sometimes regarded as a secondary cause of situation or change of state: przeszyty włócznią 'speared with a spear', lina$_\theta$ cumownicza$_{adjective}$ 'a hawser, lit. mooring rope$_\theta$'.

Material$_\theta$ is an entity that is used by someone to produce something from it, material undergoes change of state resulting in its disappearance and emerging of a result: zrobiony z mosiądzu$_\theta$ 'made out of brass$_\theta$', mosiężna$_\theta$ figurka 'brass$_\theta$ statuette'.

Purpose$_\theta$ – an entity or a situation toward which the event is directed or an individual which benefits from the event (purpose combines goal, beneficiary and recipient roles): wręczenie (medali) olimpijczykom 'giving (medals) to Olympians$_\theta$, sala koncertowa$_\theta$ 'a concert$_\theta$ hall'.

Location is a physical place at which a given event is localised, a place being destination of an event, a path or a source of motion, or simply a place at which a particular individual is situated: wręczenie (medali) w auli$_\varrho$ 'giving (medals) at the lecture theatre$_\varrho$', przedzieranie się przez moczary$_\varrho$ 'struggling through the swamp$_\varrho$'.

Time is a particular moment or a duration of an event – it localises a situation within the flow of events or gives its duration: przedzieranie się przez godzinę$_\varrho$/w środę$_\varrho$ 'struggling for an hour$_\varrho$/on Wednesday$_\varrho$'.

Temporal/spatial meronymy – these relations point onto a spatial or temporal part of a place/location/time/period): poniedziałkowy poranek$_\varrho$ 'Monday morning$_\varrho$', środek$_\varrho$ zimy 'middle$_\varrho$ of the winter', koniec$_\varrho$ drogi 'end$_\varrho$ of the road', stolica$_\varrho$ kraju 'capital$_\varrho$ of the country'.

Attribute is a property of an individual or an event, such as colour, size, weigth, intensity, duration etc., which might be expressed with a qualitative adjective: czerwony$_\varrho$ samochód 'red$_\varrho$ car', głośna$_\varrho$ muzyka 'loud$_\varrho$ music'.

Family (member) is a relative or an in-law to someone, the relation is bidirectional and reflexive: syn$_\varrho$ króla$_\varrho$ 'king's$_\varrho$ son$_\varrho$', moja$_\varrho$ żona$_\varrho$ 'my$_\varrho$ wife$_\varrho$' (I am a relative to my wife).

Order gives a position of an entity or an event in an ordered sequence/chain: druga$_\varrho$ odpowiedź '2nd answer', lata 80_ϱ. 'eighties, lit. eightieth$_\varrho$ years'.

Quantity is an amount of something or a cardinality of a given set: pięciu$_\varrho$ panów 'five$_\varrho$ men',

[2] In Polish, as in other Indo-European languages, verbs could be nominalized during a process of syntactic transformation (Jędrzejko, 1993), (Kolln, 1990). Such nominalized predicates could be linked with nouns by thematic relations.

[3] Rationale for selection of the presented semantic relation types was their frequencies in a four-text sample taken from a Polish corpus *KPWr*. Together chosen relations account for ca 80% of all semantic relation occurrences in these texts. Most of our relation types could be found on the list of the most frequent relation types in the English noun phrase (Moldovan et al., 2004, Tab. 1).

kieliszek$_{\varrho}$ wina 'glass$_{\varrho}$ of wine'.

4 Semantic relation recognition rule-based algorithm

Our rule-based system proceeds in two steps[4]: first it recognizes word pairs and triples, then operators classifying relations enter.

4.1 Recognizing word pairs and triples

Since we consider relations within noun phrases, we must identify them correctly. We made use of a CRF shallow parser (Radziszewski and Pawlaczek, 2012) trained on an annotated corpus of Polish (KPWr) (Broda et al., 2012) which comprises shallow syntactic annotation level (Radziszewski et al., 2012).

KPWr contains 326 annotated text samples representing different genres and styles: blogs, press articles, official and legal texts and Polish Wikipedia articles, it comprises 106358 annotations (phrases and phrase heads, and predicate-argument relations).

Noun and preposition phrases (NPs/PPs) from the corpus correspond to arguments of predicate-argument structure. Each such NP/PP constists of one or several smaller phrases based on agreement (*AgPs*, for details, please look at cited works). Here is an example NP from the corpus (a head of the phrase is boldfaced, AgP heads are underlined):

> [[**samolot** wyprodukowany]$_{AgP}$ [przez PZL]$_{AgP}$ [w roku 1938]$_{AgP}$ [w Łodzi]$_{AgP}$]$_{NP}$
>
> 'aircraft made by PZL in (year) 1938 in Łódź (city)'

There is no reliable deep parser for Polish (Gołuchowski and Przepiórkowski, 2012), thus we decided to construct a simple rule-based algorithm for deepened shallow parsing of Polish NPs/PPs. The algorithm works on tagged texts – we used (Radziszewski, 2013) tagger. Parsing rules make use of an output from the CRF shallow parser (Radziszewski and Pawlaczek, 2012), in particular: borders of whole NPs/PPs, and of their constituents (i.e., phrases based on agreement, *AgPs*). Found pairs and triples are directly connected within a syntactic structure.

Hand-written rules act like a partial dependency parser. The pairs consist of one subordinate and one superordinate token, the triples comprise one superordinate token and a subordinate preposition phrase (preposition + governed nominal head of a subordinate noun phrase).

The whole algorithm runs in a main loop which iterates AgP$_i$ heads. We start from the first AgP$_0$ head to the left, then we proceed to the right, jumping from AgP$_i$ head to the closest AgP$_{i+1}$ head to the right. For every AgP$_i$ head we run a cascade-like chain of rules (numbered from 1 to 7) for genetives, nominatives, small preposition phrases (being a part of larger NPs or PPs), coordination, other known to the tagger tokens, other unknown to the tagger tokens and for modifiers. The algorithm in pseudocode was shown in Algorithm 1

The algorithm gives following description for just analysed phrase, "R + number" denotes the number of a rule in the Algorithm 1 activated on the word pair or triple (for instance, *R3* means that the rule number 3 was activated): *R7*: samolot ← wyprodukowany 'plane made', *R3*: wyprodukowany ← przez PZL 'by PZL',*R3*: wyprodukowany ← w roku 'in year', *R3*: wyprodukowany ← w Łodzi 'in Łódź' .

Such simple shallow parsing algorithm operates quite well on an annotated part of KPWr with F-measure equal to 84%, P = 88%, R = 80%.[5]

4.2 Applying WCCL operators

Having identified pairs and triples we run on them operators written in a constraint language WCCL (Radziszewski et al., 2011). The operators are language-specific and utilize morphosyntactic features (POS, case, number and gender), domains of Polish WordNet lexical units (word-sense pairs (Maziarz et al., 2012)), thousands of derivational relation instances between nouns, adjectives and verbs from the wordnet[6] and information about syntactic frames of nominalized predicates, taken from Polish valence dictionary (Dębowski and Woliński, 2007).

Each of written operators refers to one semantic relation. In other words, each semantic relation is described by one or by many WCCL operators. If an operator is successfully applied to a pair (or a

[4]Similarly to system presented in (Gamallo et al., 2002).

[5]Random sample of 200 NPs/PPs taken from KPWr, 331 relation instances, bootstrap confidence intervals are following P = 83÷91%, R = 76÷84%, F = 79÷87%, $\alpha = 0.05$. The corpus was divided by us into two parts: one working set for testing and preparing parsing rules and semantic operators - consisting of 300 texts, and a smaller evaluation part of 26 texts.

[6]Since we do not use any word sense disambiguation system, we simply take the first sense of every given word.

Algorithm 1 Rule-based algorithm for the recognition of word pairs and triples

1. **genetive attachment** – link AgP_i head in genetive to the closest AgP_{i-1} head to the left or to the closest nominalized predicate to the left:
   - if there is none - link it to the closest predicate to the right;
   - if there is none - link the considered AgP_i head to the head of the whole NP/PP;
2. **nominative attachment** – link AgP_i head in nominative to the closest AgP_{i-1} head to the left or to the closest nominalized predicate to the left:
   - if there is none - link it to the closest AgP_{i+1} head to the right or to the closest nominalized predicate to the right;
   - if there is none - link the considered AgP_i head to the head of the whole NP/PP;
3. **small PP attachment** – link a head of AgP_i containing a small PP to the closest nominalized predicate to the left:
   - if there is none - to the closest nominalized predicate to the right such that it is not an element of $\text{AgP}_{j>i}$ containing a preposition;
   - if there is none - to the closest AgP_{i-1} head to the left;
   - if there is none - link AgP_i with our whole NP/PP head;
4. **coordinated syntactic groups** – look for such AgP_i that is preceded by a coordination conjunction (i.e., *i* 'and', *oraz* 'and', *lub* 'or') or by coordinating comma ('coordinating comma' is such a comma that is placed between two AgPs whose heads are agreed on case), such coordination marker cannot be an element of any AgP:
   - if there is such a marker, look to the left in order to find such $\text{AgP}_{j<i}$ head which is agreed on case with our AgP_i head – then create a new relation instance by copying the link $\text{AgP}_j \rightarrow X$ and replacing AgP_j head by the AgP_i head in that copied linkage, i.e., create the relation instance $\text{AgP}_i \rightarrow X$;
   - if it is not possible – do not introduce any relation;
5. head token provided with **POS known to the CRF tagger** – link the AgP_i head to the closest nominalized predicate to the left:
   - if there is none - to the closest nominalized predicate to the right such that it is not an element of $\text{AgP}_{j>i}$ containing a preposition;
   - if there is none - link AgP_i head to the closest $\text{AgP}_{j<i}$ head to the left such that $\text{AgP}_{j<i}$ does not contain any preposition;
   - if there is none such $\text{AgP}_{j<i}$ – connect AgP_i to the whole NP/PP head;
6. **other cases** (the AgP_i head was not provided any known POS by the CRF tagger) – in such cases link AgP_i head to the closest $\text{AgP}_{j<i}$ head to the left; if there is none – do not make any decision;
7. **relations inside AgPs** – link adjectival and participial modifiers to the head of AgP_i.

triple), then we know what semantic relation between the pair (or triple) occurs. Otherwise, we assume that the semantic relation does not occur.

For example, our `Proto-Patient` relation was described by the 6 WCCL operators. One of them is presented in Listing 1. This operator uses two dictionaries with valence frames (`acc` - a list of verbs possessing any accusative frame, `frames` - a list of verbs described in the Polish valence dictionary (Dębowski, 2013)) and morphosyntactic information about part of speech (`class`) and `case`.

This operator `PROTO-PATIENT-acc` captures pairs like *dręczący*$_{pact}$ *Janka*$_{noun.acc-\theta}$ 'tormenting John$_\theta$' with a noun playing a Proto-Patient role of the predicate *dręczący*. The operator first checks whether a predicate (active participle) has an accusative frame or is outside the dictionary of Dębowski (`"frames"`). Since *dręczyć* 'to torment' is in `acc` dictionary and since *Janek* 'John' has `subst` class and `acc` case - the boolean operator returns 'true'.

Let us present another example: the Proto-Agent macrorole is recognized by 5 operators, in Listing 2 was shown one of them. The `PROTO-AGENT-ger-przez-acc` operator is written for triples, i.e., for a triple *wydanie*$_{pact}$ *przez*$_{pron}$ *wydawcę*$_{noun.acc-\theta}$ 'publishing by the

publisher$_\theta$'. The first element in the triple is a gerund form of verb *wydać* 'to publish'. The operator checks whether the verb *wydać* has in its frame accusative/genetive or whether it cannot be found in Dębowski's dictionary (position 0 in the triple, `frames`).

Listing 1: One of the WCCL operators describing Proto-Patient relation. Language details has been described in (Radziszewski et al., 2011), abbreviations for grammatical categories has been explained in (Przepiórkowski et al., 2012)

```
@b:"PROTO-PATIENT-acc" (
 and(
  // 0 - accusative frame
  equal(class[0], pact),
  or(
   equal(lex(base[0], "acc"), ["1"]),
   not(equal(
   lex(base[0], "frames"), ["1"]))
  ),
  // 1 - noun or adj. & accusative
  in(class[1], {subst,depr,ger,adj}),
  equal(cas[1],acc)
 )
)
```

Next the operator seeks for the preposition *przez* 'by' at position 1. Then it tests if the first meaning of the lemma *wydawca* 'publisher' does not belong to the domain 'time' (= Polish `czas`) in Polish WordNet (position 2). Indeed, the first meaning of *wydawca* is in the domain 'person' (that iformation is avaiable in the dictionary `noun_domain`). At the end, we check whether the last token of our triple is in accusative. Because all of these conditions are fulfilled, the operator returns 'true', and we may assume that the last token takes the role of Proto-Agent.

Listing 2: A WCCL operator for the Proto-Agent relation

```
@b:"PROTO-AGENT-ger-przez-acc" (
 and(
  // 0 - gerund
  equal(class[0],{ger}),
  or(
   equal(lex(base[0], "acc"), ["1"]),
   equal(lex(base[0], "gen"), ["1"]),
   not(equal(
    lex(base[0], "frames"), ["1"]))
  ),
  // 1 - preposition "przez"
  equal(orth[1],"przez"),
  // 2 - not 'time' & accusative
  equal(cas[2], acc),
  not(
   equal(lex(if(
    equal(class[2], {ger}),
    lex(base[2], "ger_base"), base[2]),
   "noun_domain"), ["czas"]))
 )
)
```

In Listing 3 one operator for family ralation was shown. `FAMILY-agpp` used to recognize this relation for word pairs. The operator, inter alia, uses semantic dictionary of kinship names built on the basis of Polish WordNet (the dictionary `kinship`), lammas of possessive pronouns (e.g., `mój` 'my', `twój` 'yours').

Listing 3: Two WCCL operators describing Family relation

```
@b:"FAMILY-agpp" (
 and(
  // agreement
  agrpp(0,1, {nmb, gen, cas}),
  // position 0
  in(base[0], ["moj", "twoj",
   "swoj", "nasz", "wasz"])
  // position 1
  equal(lex(base[1], "kinship"), ["1"]),
  equal(lex(
   base[1], "noun_domain"), ["os"]),
  in(class[1], {ger, subst, depr}),
 )
)
```

5 Results and conclusions

Evaluation of the presented semantic relation recognition algorithm was performed in three steps. First experiment (labelled `kpwr`) was performed on a random sample of the KPWr corpus (26 out of 326 texts, aproximately one thirteenth of the corpus). In this experiment we made use of syntactic annotations from KPWr (cf. Tab. 1). Second experiment was performed on a random sample of 100 texts taken from yet another Polish corpus, called *NKJP* (Przepiórkowski et al., 2012, `nkjp`, approximately one tenth of the corpus)[7]. Since NKJP lacked syntactic annotations of KPWr style, we were forced to run on it the CRF shallow parser (described in Sec. 4.1). This experiment gave us information about performance of our algorithm on a 'bare' text (see Tab. 2). Evaluation in the experiments was done by a professional linguist.

At last, four baseline models were constructed and evaluated on the two corpora (Tab. 3). We created baselines similar to that presented in (Uchiyama et al., 2008), which was majority model. We chose the most frequent relation, which in the sample from KPWr was Proto-Patient (with the number of 113 instances out of 268 relation instances), this relation type was also the most frequent in the sample of NKJP (411 out of 1950 relation instances). For each corpora two baselines

[7] We focused on one-million balanced version of the much bigger corpus.

Relation	TP/FP/FN	P [%]	R [%]	F1 [%]
Proto-Agent	7/5/16	58.3	30.4	40.0
Proto-Patient	45/8/68	84.9	39.8	54.2
Instrument	0/0/7	—	0.0	—
Material	0/0/3	—	0.0	—
Purpose	1/7/30	12.5	3.2	5.1
location	3/9/25	25.0	10.7	15.0
sp. meronymy	0/3/2	0.0	0.0	—
time	2/2/3	50.0	40.0	44.4
t. meronymy	1/0/1	—	—	—
attribute	14/18/10	43.8	58.3	50.0
family	0/0/2	—	—	—
order	5/0/5	100.0	50.0	66.7
quantity	10/2/8	83.3	55.6	66.7
All	88/54/186	53.8-70.1	*26.8-38.0	*36.2-48.3

Table 1: Results of the algorithm on a sample from KPWr: P = Precision, R = recall, F1 = F-measure, TP = true positives, FP = false positives, FN = false negatives, sp. = spatial, t. = temporal. Percentile bootstrap confidence intervals are calculated at $\alpha = 0.05$. Asterisks denote significant differences between `kpwr` and `nkjp` in one-tailed tests, $\alpha = 0.05$

Relation	TP/FP/FN	P [%]	R [%]	F1 [%]
Proto-Agent	75/7/143	91.5	34.4	50.0
Proto-Patient	181/17/230	91.4	44.0	59.4
Instrument	2/1/8	66.7	20.0	30.8
Material	3/4/36	42.9	7.7	13.0
Purpose	13/7/94	65.0	12.2	20.5
location	90/75/202	54.6	30.8	39.4
sp. meronymy	12/11/25	52.2	32.4	40.0
time	25/16/75	61.0	25.0	35.5
t. meronymy	2/0/66	100	2.9	57.1
attribute	200/248/64	44.6	75.8	56.2
family	18/0/6	100.0	60.0	85.7
order	33/0/100	100.0	24.8	39.8
quantity	113/68/146	62.4	43.6	51.4
All	767/454/1195	60.1-65.6	*36.9-41.2	*46.0-50.3

Table 2: Results of the algorithm on a sample from NKJP, labels as in the previous table. Percentile bootstrap confidence intervals are calculated at $\alpha = 0.05$. Asterisks denote significant differences between `kpwr` and `nkjp` in one-tailed tests, $\alpha = 0.05$

were calculated: in Baseline #1 we assumed that we had perfectly recognized all occurences of semantic relations (of any type), in Baseline #2 we simply signed with 'Proto-Patient' label every recognized by our system semantic relation instance. Baseline #2 is realistic, while #1 is idealistic, since to obtain #1 we should be able to recognize every single relation instance within a corpus. Baselines #1 are upper limits for all majority models (including #2). Our two idealistic baselines are higher than the realistic baselines (see Tab. 3).

Percentile bootstrap methods (DiCiccio and Efron, 1996), (DiCiccio and Romano, 1988) were applied to statistical significance and confidence interval (*CI*) analysis of the data.[8] We took 10000

[8]Our data for NKJP were merged, so cross-validation was not avaiable.

`kpwr`	P	R	F1
Baseline #1	*42.2%	***42.2%**	**42.2%**
Baseline #2	*26.2%	*20.0%	*22.7%
Experiment	**62.0%**	32.8%	**42.9%**
`nkjp`	P	R	F1
Baseline #1	*21.1%	*21.0%	*21.0%
Baseline #2	*14.9%	*9.2%	*11.4%
Experiment	**62.5%**	**38.9%**	**47.9%**

Table 3: Precision, recall and F1 for baselines (#1 & #2) and experiments (`kpwr`, `nkjp`). Asterisks denote significant differences between an experiment and a baseline in one-tailed test at $\alpha = 0.05$

bootstrap resamplings for each measure (P, R, F1), α was equal to 0.05 for each one-tailed test and CI (a percentile CI need not be symmetrical).

In `nkjp` we have beaten both idealistic and realistic baselines. Precision, recall and F1 for `kpwr` are higher than Baseline #2. Only idealistic Baseline #1 for the KPWr corpus has overtaken our rule-based algorithm with regard to recall (42.2% vs. 32.8%), while its precision is lower and F1's are statistically indistinguishable.

Results are promising, precisions go above 50% (the lower endpoint for the `kpwr` confidence intervel), for `nkjp` we may assess it even more precisely as 60%-65%. Some semantic relations are recognized with higher precision: Proto-Agent (`nkjp`: 89-100%, `kpwr`: 90-100%, $\alpha = 0.05$), Proto-Patient (`nkjp`: 88-95%, `kpwr`: 83%-98%), family (`nkjp`: 90-100%) and order (`nkjp`: 91-100%). Our system is thus comparable in this aspect to the systems described in Sec. 2.[9]

Overall recall is low, but higher than realistic baselines. In `kpwr` we obtained R = 27-38%, while for `nkjp` we got statistically higher interval of 37-41%. It seems that recall was not affected by lack of marked NP/PP borders in the corpus (these should have been brought out by the CRF shallow parser). F-measures calculated on our both corpora are also much higher than realistic baselines #2.

We can already conclude that our preliminary experiments turned successful. Now we are aiming at improving our operators to raise their recall and at expanding the semantic role set (e.g., for Agent, Causer, Experiencer, Possessor or Result). Parallel, we start work on construction of automatic algorithms for relation recognition.

[9]Not directly, of course.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(5):1–11.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*, Istanbul, Turkey. ELRA.

Thomas J. DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212.

Thomas J. DiCiccio and Joseph P. Romano. 1988. A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):338–354.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Łukasz Dębowski and Marcin Woliński. 2007. Argument co-occurrence matrix as a description of verb valence. In *Proc. of the 3rd Language & Technology Conference*, volume 3, pages 260–264, Poznań, Poland.

Łukasz Dębowski. 2013. Polish valence dictionary (http://nlp.ipipan.waw.pl/ppjp/ slownik/swigra/koncowy.txt).

Pablo Gamallo, Marco Gonzalez, Alexandre Agustini, Gabriel Lopes, and Vera S. De Lima. 2002. Mapping syntactic dependencies onto semantic relations. In *ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*.

Konrad Gołuchowski and Adam Przepiórkowski. 2012. Semantic role labelling without deep syntactic parsing. In Hitoshi Isahara and Kyoko Kanzaki, editors, *Advances in Natural Language Processing*, volume 7614 of *Lecture Notes in Computer Science*, pages 192–197. Springer Berlin Heidelberg.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rongfeng Huang. 2009. Semantic relation extraction by automatically constructed rules. In *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, AICI '09, pages 425–434, Berlin, Heidelberg. Springer-Verlag.

Ewa Jędrzejko. 1993. *Nominalizacje w systemie i tekstach współczesnej polszczyzny*. Uniwersytet Śląski, Katowice.

Kate Kearns. 2011. *Semantics*. Palgrave.

Martha J. Kolln. 1990. *WordNet: Understanding English Grammar*. Macmillan.

Gabriel Larson, Richard & Segal. 1996. *Knowledge of Meaning: An Introduction to Semantic Theory*. The MIT Press.

Roman Laskowski and Henryk Wróbel, editors. 1997. *Gramatyka współczesnego języka polskiego. Morfologia [Grammar of contemporary Polish: Morphology]*. PWN.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November.

2009. Luna project: Spoken language understanding in multilingual communication systems (http://zil.ipipan.waw.pl/luna).

Marek Maziarz, Adam Radziszewski, and Jan Wieczorek. 2011. Chunking of Polish: guidelines, discussion and experiments with Machine Learning. In *Proc. of the LTC*.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan, January.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 60–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vivi Nastase, Jelber Sayyad Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

Palmer, Martha & Gilldea, Daniel & Xue, and Nianwen. 2010. *Semantic Role Labeling*. Morgan & Claypool Publishers.

Adam Przepiórkowski, Miroslaw Banko, Rafal Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Jezyka Polskiego*. Wydawnictwo Naukowe PWN.

Adam Radziszewski and Adam Pawlaczek. 2012. Large-scale experiments with NP chunking of Polish. In *Proceedings of TSD 2012*, Brno, Czech Republic. Springer.

Adam Radziszewski, Adam Wardyński, and Tomasz Śniatowski. 2011. WCCL: A Morpho-syntactic Feature Toolkit. In *Proceedings of the Balto-Slavonic Natural Language Processing Workshop*. Springer.

Adam Radziszewski, Marek Maziarz, and Jan Wieczorek. 2012. Shallow syntactic annotation in the Corpus of Wrocław University of Technolog. *Cognitive Studies*, 12.

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 256–259, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01)*, pages 82–90.

Stephen Tratz and Eduard Hovy. 2010. Isi: Automatic classification of relations between nominals using a maximum entropy classifier. In *Proc. of the 5th Intern. Workshop on Semantic Evaluation*, Sweden. Association for Computational Linguistics.

Kateryna Tymoshenko and Claudio Giuliano. 2010. Fbk-irst: Semantic relation extraction using cyc. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 214–217, Uppsala, Sweden, July. Association for Computational Linguistics.

Kiyoko Uchiyama, Shunsuke Aihara, and Shun Ishizaki. 2008. Identifying semantic relations in japanese compound nouns for patent documents analysis. In *Proceedings of the 3rd international conference on Large-scale knowledge resources: construction and application*, LKR'08, pages 75–81, Berlin, Heidelberg. Springer-Verlag.

Robert D. Van Valin. 2004. Semantic macroroles in role and reference grammar. In Rolf Kailuweit and Martin Hummel, editors, *Semantische Rollen*, pages 62–82. Tuebingen Narr.