INTERSPEECH 2009 BRIGHTON

Large-Scale Polish SLU

Patrick Lehnen[1], *Stefan Hahn*[1], *Hermann Ney*[1], *Agnieszka Mykowiecka*[2]

[1]Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Germany.
{lehnen, hahn, ney}@cs.rwth-aachen.de
[2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland and
Polish-Japanese Institute of Information Technology, Warsaw, Poland.
agn@ipipan.waw.pl

Abstract

In this paper, we present state-of-the art concept tagging results on a new corpus for Polish SLU. For this language, it is the first large-scale corpus (∼200 different concepts) which has been semantically annotated and will be made publicly available. Conditional Random Fields have proven to lead to best results for string-to-string translation problems. Using this approach, we achieve a concept error rate of 22.6% on an evaluation corpus. To additionally extract attribute values, a combination of a statistical and a rule-based approach is used leading to a CER of 30.2%.

Index Terms: Polish, spoken language understanding, conditional random fields, tagging

10.21437/Interspeech.2009-696

1. Introduction

Spoken language understanding (SLU) is a well-known field of research concerning machine learning. Only in recent years, larger scale corpora collections for Polish have started, e.g. [1]. Unfortunately there are still very few speech corpora and they are not semantically annotated (cf. [2, 3]). Such corpora would allow us to use state-of-the-art data driven machine learning approaches.

In this paper, we present our recent experiments on the task of concept tagging using a newly created semantically annotated corpus in the domain of transportation information [4]. Concept tagging is usually defined as the segmentation and labeling of a given word sequence into smallest units of meaning, which may be task dependent. Additionally to this segmentation, so-called attribute values may be extracted from the segments, which reflect the most important information w.r.t. the concept. An example from the corpus is given in Figure 1. At the top, the original spoken word sequence is given, followed by the English translation to facilitate understanding. Afterwards, the XML-annotation is presented. Each concept is represented with one line starting with the chunk ID, followed by the word span, the attribute name and the corresponding attribute value.

In recent years, conditional random fields (CRFs) have attracted growing interest in the SLU community due to the closed mathematical framework and their properties [5]. Also for string-to-string translation tasks like transliteration or attribute name extraction, statistical models based on CRFs lead to state-of-the art results [6, 7]. Thus, they are an effective approach to solve the tasks relevant for this paper, i.e. attribute name and value extraction.

The following section presents task and corpus chosen for this paper. Afterwards, the particularities of the Polish language which have to be kept in mind when dealing with statistical approaches to concept tagging are presented. In Section 3, our approach to tackle this task, namely conditional random fields, is presented. In the following two sections 4 and 5 our approaches to attribute name and value extraction are described. Section 6 presents our experimental findings. A conclusion is given in Section 7, which is followed by an outlook.

[. . .] chciałam linię sto pięćdziesiąt jeden [. . .] z Płockiej
w kierunku Centrum w dni świąteczne
[. . .] I would like line one hundred fifty one [. . .] from Płocka
in the direction of Centrum on holidays
< [. . .] />
<c_id="4" span="word_" attr="Action" value="Request" />
<c_id="5" span="word_10..word_13" attr="BUS" value="151" />
value="Młynów" />
< [. . .] />
<c_id="8" span="word_19..word_20" attr="SOURCE_STR" value"Płocka" />
<c_id="9" span="word_21..word_23" attr="GOAL_DIRECTION_TD"
value="Centrum" />
<c_id="10" span="word_24..word_26" attr="TIME_PERIOD"
value="Holiday" />

Figure 1: An example of a word sequence annotated with attribute names and values. To facilitate understanding, the English translation is also given.

2. Corpus

The corpus of transportation information dialogues collected under the LUNA project is the first Polish corpus with semantically annotated speech data which will be publicly available (the corpus will be available together with its description in [8]).

2.1. Task Description

The chosen application domain is public city transport network, i.e. information about stops, routes, connections, timetables and fares. Possibility of getting this information using mobile while being on a street is a quite popular service. An extension of human operated service by a similar automatic one would lower its costs and ease the access. An important element of such a system would be automatic recognition of concepts addressed in user queries. To achieve this goal a domain ontology and a model for recognizing phrases which are natural language equivalents of concepts from the ontology had to be build.

The domain of public transport related information seeking dialogs contains several important subdomains: elements of Warsaw topology (streets, squares, important buildings, etc.), public transport network description (e.g. names of lines and stops, timetables, etc.), temporary changes of traffic rules (e.g.

Table 1: Quantitative characteristics of the collected corpus.

Category	# calls	avg. # user's words per call	vocabulary
Routes	93	98	1975
Itinerary	140	96	2562
Schedule	111	61	1339
Stops	55	86	1332
Reduced fares	101	48	1735
Total	500	85	4130

- (jestem) na Polnej$_{adj,fem,loc}$/Dąbrowskiego$_{adj,masc,gen}$ *(I am) on Polna Street / Dąbrowskiego Street*
- (jadę) z Polnej$_{adj,fem,loc}$ /Dąbrowskiego$_{adj,masc,gen}$ *(I am coming) from Polna Street / Dąbrowskiego Street*
- (jadę) na Polną$_{adj,fem,acc}$ / Dąbrowskiego$_{adj,masc,gen}$ *(I am going) to Polna Street / Dąbrowskiego Street*

Figure 2: Examples for the complexity of Polish morphology.

road disturbances, detours), selected people features (age, information allowing for fee reduction) and characteristic of the request types (e.g. questions about particular type of information, confirmation requests).

For representing all types of information adequate for the domain, an ontology of 240 concepts was defined.

2.2. Acquisition

The corpus of real human-human dialogues was collected during May 2007 at the Warsaw City Transportation information center where two to four persons typically answer 200-300 calls per day (most calls last from 1 to 2 minutes) [9]. At the end, 500 dialogues were chosen for the LUNA corpus. The dialogues were divided into five thematic groups (see Table 1). A detailed description of the acquisition procedure is given in [4]. The recorded dialogues were manually transcribed and then annotated on several levels – morphological, syntactic and semantic. In this paper, we address the semantic annotation consisting of concept names and their values. It was done by hand crafted rules and then manually corrected [10]. In the corpus, 205 concepts from our ontology occurred at least once. The structure of annotation files is illustrated in Figure 1.

2.3. Data specificity

Polish is an inflectional language with a relatively free word order. Polish nouns and adjectives inflect for case (seven) and number. Adjectives inflect for gender (up to five forms in singular and two in plural) and agree in case, number and gender with nouns they modify. A role of a noun phrase in a sentence is defined by its case and by a preposition (if it is present), prepositions can introduce noun phrases in one or two cases. Polish multiword proper names inflect, but not in the same way as their elements taken in isolation. In particular, names in genitive form (a typical form for streets named after people) do not inflect at all. Figure 2 shows examples of different ways of inflection for Polish location names.

The next problem which makes our task more difficult is a great number of concepts which are closely connected. For example, there are about 50 concepts describing various places and several concepts describing time. In the phrases in Figure 2, we have three different concepts describing places: LOCATION_STR, SOURCE_STR and GOAL_STR (STR is an abbreviations for street).

2.4. Partitioning of the corpus

The roughly 12.5k collected and annotated human-human dialogue turns have been split into three subsets. As preprocessing, all turns with big overlaps between operator and user have been discarded. In a first step, the operator turns have been separated from the user turns. The operator turns have a different distribution of concepts since these are often questions back to the user. Thus, they are not suitable for development or evaluation and will be used for training of the statistical models only. Due to the pretty large number of concepts in this corpus (for comparison, within the well-known MEDIA corpus, there are roughly 100 different concepts [11]), the development and evaluation sets should not be too small to avoid high OOV ratios which are also not suitable to measure performance gains in a consistent way. The user turns have been split into three sets of roughly 2k sentences. One of these parts has been added to the training set, a second part serves as development corpus, and the third and last part forms the evaluation set. Since it may be interesting for future experiments to have the dialogue context, the corpora have been arranged in such a way that all user turns of a dialogue are in exactly one of the three sets. The statistics for the resulting corpora are presented in Table 2. The number of NULL tokens refers on concept level to the number of times the "garbage" concept occurs in the respective part. This concept represents chunks of the turn without any semantic meaning relevant for the task. The number of NULL words is calculated by counting all words which are tagged with the NULL concept. These figures can in some way be compared to the silence ratio known from speech recognition, since this material does not contain information to discriminate between the interesting classes.

3. Conditional Random Fields

Conditional Random Fields are discriminative log-linear models describing the probability of a sequence of output words c_1^N based on a given sequence of input words w_1^N [5]. They are normalized on the target sentence level c_1^N and defined by a large set of real valued feature parameters λ_m. In CRFs, the feature functions $h_m(c_{n-1}, c_n, w_{n-2}^{n+2})$ are providing the degrees of freedom to describe the training material $\{(c_1^N|w_1^N)\}_1^T$. In general, these feature functions do not need to be orthogonal.

Linear Chain Conditional Random Fields (CRFs) as defined in [5] are special CRFs expecting the output sentence c_1^N to be ordered as a linear chain. They can be represented with equation 1:

$$p(c_1^N|w_1^N) = \frac{1}{Z}\prod_{n=1}^{N}\exp\left(\sum_{m=1}^{M}\lambda_m h_m(c_{n-1}, c_n, w_1^N)\right) \quad (1)$$

using

$$Z = \sum_{\tilde{c}_1^N}\prod_{n=1}^{N}\exp\left(\sum_{m=1}^{M}\lambda_m h_m(c_{n-1}, c_n, w_1^N)\right). \quad (2)$$

Most publications describing applications of CRFs actually use linear chain CRFs.

In our experiments we use binary feature functions $h_m(c_{n-1}, c_n, w_1^N) \in \{0, 1\}$. If a pre-defined combination of

Table 2: Statistics of the Polish training, development and evaluation corpora.

corpus	training		development		evaluation	
POLISH-SLU	words	concepts	words	concepts	words	concepts
# sentences	8,341		2,053		2,081	
# tokens	53,418	28,157	13,405	7,160	13,806	7,490
# NULL tokens	21,973	9,811	5,680	2,384	5,743	2,486
vocabulary	4,081	195	2,028	157	2,057	159
# singletons	1,818	19	1,119	23	1,113	28
# OOV rate [%]	–	–	4.95	0.13	4.96	0.11

the values $c_{n-1}, c_n, w_{n-2}, \ldots, w_{n+2}$ is found within the data, the value "1" is returned, otherwise the value "0". E.g., a feature function may fire if and only if the predecessor word w_{n-1} is "the" and the concept c_n is "name". We apply feature functions based on predecessor, the current, and successor words (*lexical features*), features based on the predecessor concept (*transition features*) and *word part features* capturing pre- and suffixes as well as capitalization.

On a given training dataset $\{\{c_1^N\}_t, \{w_1^N\}_t\}_{t=1}^T$, the feature parameters λ_m are estimated by an iterative optimization of the conditional log-likelihood using a regularization prior $c|\lambda_1^M|^2$ with a regularization parameter c, while the decision is based on the maximization of the probability $p(c_1^N|w_1^N)$.

4. Attribute Name Extraction

Starting from the input sentence, e.g. from a phone call requesting for a timetable information, the first processing step is to find the location of content words in the input sentence and assign the corresponding attribute names, e.g. the bus number.

@Action{chciałam} @BUS{linię sto pięćdziesiąt jeden} ...
@Action{I would like} @BUS{line one hundred fifty one} ...

Since our modeling approach relies on a 1-to-1 mapping between word and attribute name sequence, the attribute names are usually broken down in start and continue tags, e.g *start_bus*. Thus, it is ensured that the word sequence has the same length as the attribute name sequence.

start_Action:chciałam start_bus:linię bus:sto bus:pięćdziesiąt bus:jeden ...

In general during search CRFs permit an attribute name tag sequence `start_A` **`A B`**, which can not be seen in training, since it conflicts with the *start tag* rule. This problem can be solved by either interpreting a transition `A→B` as `A→start_B` or reducing the search space by all conflicting transitions like `A→B`. In our experiments we always obtain better results by interpreting a transition `A→B` as `A→start_B`.

5. Attribute Value Extraction

Knowing the location and the attribute name of content words given by the attribute name extraction, the next step is to extract normalized values for most of the attribute names, e.g. concerning the example from the previous subsection *Request* or *151*:

@Action[Request]{chciałam} @BUS[151]{linię sto pięćdziesiąt jeden} ...
@Action[Request]{I would like} @BUS[151]{line one hundred fifty one} ...

The number of possible values varies highly between attribute names. For example, the attribute name *Reaction* can take either the value "Confirmation" or "Negation" and is triggered by only few content words. In contrast, the value of *STREET_NUMB* can at least theoretically be any number. In principle, attribute value extraction can be realized using machine learning. This is a quite easy task when the number of possible values is low but can become difficult for attribute names with a huge number of possible values like street or bus numbers. These numbers can not be covered completely by the training corpus, which is the only information source at least for purely data driven approaches. A 1-to-1 mapping like in attribute name extraction is not used, instead exactly one value is hypothesized per attribute name. As features, lexical features on the predecessor, the current, and the successor word were used. For attribute names with a huge number of values, it is possible to reduce the search space only to a *null* value, leaving the attribute value extraction to a rule based approach in a possible post-processing step. In the experiments described the rule-based attribute value extraction has been applied to the seven most error-prone attribute names.

6. Experimental Results

The performance of our models has been measured using the well-known concept error rate (CER) as metric. If attribute values are extracted additionally, the concept together with the value has to be correct to not lead to an error. The evaluation is done using the NIST toolkit [12]. In this section, we will first describe the optimization process and feature selection for our Polish tagging system. Using a starting lexical window size of a width of one around the current word, i.e. $[-1, .., 1]$, the regularization parameter c (L2 regularization term on the CRF parameters) is tuned. The best performance (25.7% CER on DEV) has been achieved using $c = 1/64$. This system sets our baseline. Afterwards, word part features are introduced and optimized. These features include prefix, suffix and capitalization features. Since prefix features lead to the largest gain w.r.t. the word part features, they are tuned next. The size of the prefix window is enlarged continuously (including the prefixes of smaller size), until an optimum is found on the DEV corpus. We get an improvement of approx. 11% relatively down to 22.8% CER. Since Polish inflection changes mainly suffixes, we expect that the prefix features cover mainly the word stem. Suffix features are introduced and tuned in the same manner, leading to an additional reduction in CER of approx. 3% down to 22.0%. Adding the capitalization features gives a marginal improvement, but since it is cheap w.r.t. computational time, it is also included. The final system has an error rate of 22.0% on the DEV corpus and 22.6% on the EVA corpus. An overview of these results is given in Table 3.

Assigning 200 concepts had to lead to a big variety of errors. As was expected, there are quite a lot of errors with assigning highly related concepts. For example, for 13 concepts representing goals (GOAL_X concepts where X stands for different types of locations, i.e. streets, buildings, areas, etc.) which oc-

Table 3: Concept Error Rates (CER) for various feature settings on the Polish DEV and EVA corpora. For attribute name and value extraction, results are also given for the reference concept sequence.

	CER [%] name		CER [%] name+value	
features [window]	DEV	EVA	DEV	EVA
lexical [-1..1] + trans.[-1]	25.7	26.1	-	-
+prefixes [1..4]	22.8	23.5	-	-
+suffixes [1..4]	22.0	22.7	-	-
+capitalization	21.8	22.6	31.6	32.1
+ attr. value rules	21.8	22.6	30.1	30.2
reference	-	-	16.8	17.2
+ attribute value rules	-	-	14.7	14.6

curred 275 times in EVA corpus, there were 78 errors including 58 substitutions of which 25 were substitutions of one GOAL subtype by another, 27 were other location concepts and only one was the completely wrong TIME_REL concept. What might have not been obvious from the beginning, is the fact that concepts which are very hard to recognize are questions. Apart form confirmation question which were recognized pretty well (12 errors for 164 concept occurrences), the other questions were recognized rather poorly (157 errors for 234 occurrences). This is probably due to fact that in the recorded dialogues typical questions are not formally constructed - they are usually only marked by pronunciation or by introducing words which can be also interpreted differently (e.g. interrogative particle 'czy' can also mean 'whether'). Second important and not well recognized concept was STREET (35 errors for 76 occurrences).

The results of concept value extraction (CER of 32.1% on evaluation corpus) shows that recognition of a short list of possible concept values using CRF is quite efficient. On the other hand, recognition of proper names was not so good and a list of all names and their types improved both value and concept type assignment. The next typical error observed was incomplete recognition of time descriptions for which only an hour part was identified. This issue was solved by addition of rules describing Polish time description (CER of 30.2% on evaluation corpus). If we assume that the manual concept annotation of the corpus is flawless, 17.2% of the attribute values are extracted wrongly using the purely statistic approach. The combination with rules leads to 14.6%. All results are presented in Table 3.

7. Conclusion

In this paper, we have presented state-of-the-art tagging results on the first large-scale corpus for Polish SLU. The corpus collection process as well as the problems originating from the complexity of the task and data specialities have been discussed. We have chosen to apply CRFs for attribute name and value extraction, whereas for the latter one we did a combination with a rule-based approach. Our final models lead to a CER of 22.6% for attribute name extraction and 30.2% if attribute values are additionally extracted.

8. Outlook

Since the original recordings of the dialogues are available, it would be interesting to produce concept tagging results on automatic transcriptions using an ASR system instead of manual transcriptions as has been done so far. Thereby, the use of word lattices may lead to improvements over single best hypotheses. Concerning the CRF model, we are investigating other features and categorization. There may also be possibilities to optimize the attribute value extraction process by a deeper error analysis and improved rules for the values where the statistical model fails.

9. Acknowledgements

This work was partly funded by the European Union under the integrated project LUNA - spoken language understanding in multilingual communication systems (FP6-033549).

10. References

[1] A. Przepiórkowski, *The IPI PAN Corpus. Preliminary version*. Warsaw, Poland: ICS PAS, 2004.

[2] K. Marasek and R. Gubrynowicz, "Multi-level annotation in SpeeCon Polish speech database," in *Intelligent Media Technology for Communicative Intelligence, 2nd Int. Workshop, IMTCI 2004, LNAI 3490*. Springer, 2005, pp. 58–67.

[3] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, M. Lange, D. Śledziński, and N. Cylwik, "Jurisdic – Polish speech database for taking dictation of legal texts," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008, pp. 1280–1287.

[4] A. Mykowiecka, K. Marasek, M. Marciniak, J. Rabiega-Wiśniewska, and R. Gubrynowicz, "Annotation of Polish spoken dialogs in LUNA project," in *3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC)*, Poznan, Poland, Oct. 2007.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, Jun. 2001, pp. 282–289.

[6] T. Deselaers, S. Hasan, O. Bender, and H. Ney, "A deep learning approach to machine transliteration," in *Proceedings of the EACL 2009 Workshop on Statistical Machine Translation*, Athens, Greece, Mar. 2009, pp. 233–241.

[7] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, "A comparison of various methods for concept tagging for spoken language understanding," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

[8] M. Marciniak, Ed., *Anotowany korpus dialogów telefonicznych*. Warsaw, Poland: EXIT, in preparation.

[9] K. Marasek and R. Gubrynowicz, "Design and data collection for spoken Polish dialogs database," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

[10] A. Mykowiecka, M. Marciniak, and K. Głowińska, "Automatic semantic annotation of Polish dialogue corpus," in *Text, Speech and Dialogue 11th International Conference, TSD 2008, Brno, Czech Republic, LNAI 5246*. Springer, 2008, pp. 625–632.

[11] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn *et al.*, "Semantic Annotation of the French Media Dialog Corpus," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 3457–3460.

[12] NIST. Speech recognition scoring toolkit (SCTK). http://www.nist.gov/speech/tools/.