Unstated Subject Identification in Czech

G. L. Nguy and M. Ševčíková
Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. In this paper we aim to automatically identify subjects, which are not expressed but nevertheless understood in Czech sentences. Our system uses the maximum entropy method to identify different types of unstated subjects and the system has been trained and tested on the Prague Dependency Treebank 2.0. The results of our experiments bring out further consideration over the suitability of the chosen corpus for our task.

Introduction

In Czech, the subject is often expressed in the surface shape of the sentence but can be omitted as well. When present, it is expressed by a sentence member (e.g. noun phrase or pronoun in nominative case, infinitive phrase) or by a dependent clause; we speak about an overt subject. Concerning sentences in which the subject position is not occupied, the subject can be understood from the respective verb form or from the context and/or situation (and added into the surface sentence structure) in most cases; these unstated (covert, null, zero, empty) subjects are the main concern of the present paper. However, there are several subjectless verbs such as *Prší* lit.: 'Rains' ('It rains'), which cannot be accompanied with a subject at all. These cases are taken into consideration in our work as well. In a subjectless finite verb clause we distinguish four following types of unstated subjects:

Implicit subject: The subject is omitted in the surface text but can be understood from the verb morphological information; most often it stands for an entity already mentioned in the text or can be deictic.

(1) Jana ráda peče. Dnes Ø upekla jablečný koláč.
Lit. Jane gladly bakes. Today [she] baked$_{3.SG.FEM}$ apple pie.
Jane likes to bake. Today she has baked an apple-pie.

General subject: The subject does not refer to any concrete entity; it has a general meaning, so it can be omitted in the surface structure.

(2) S rizikem se Ø počítá.
Lit. With risk RFLX [one] counts$_{3.SG}$.
Risk is counted in. (One counts risk in.)

Unspecified subject: The subject denotes an entity more or less known from the context which is however not explicitly referred to.

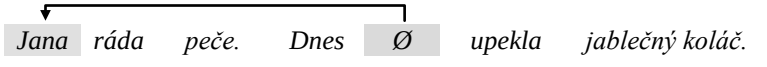
(3) Ø Hlásili to v rádiu.
Lit. [They] Announced$_{3.PL.ANIM}$ it on radio.
It was announced on radio. (They announced it on radio.)

Null subject[1]: The subject does not refer to any entity in the real world. It is neither phonetically realized, nor can be lexically retrieved. In this case the predicate is an impersonal (weather) verb.

(4) Zítra Ø bude oblačno.
Lit. Tomorrow [it] will$_{3.SG}$ cloudy.
Tomorrow it will be cloudy.

Rello & Ilisei [2009] consider another category of omitted subject, i.e. omitted subject in a non-finite verb clause. It is a case of control, but we do not study it in this paper.

In Czech, it is natural to drop out personal pronouns in subject position of the clause[2]. An overt subject pronoun indicates an emphasis of the speaker. In this paper we discuss the unstated subject identification problem, because an unstated implicit subject in third person form is often an anaphor that refers to an entity already mentioned in the text. The term 'zero pronoun identification' is used for these cases in computational approaches to anaphora resolution. An example of a zero pronominal anaphora and its possible usage in machine translation is illustrated in Fig. 1:



[1] It is a term we use in our work to be able to talk about it easily.

[2] Czech is so-called a pro-drop language. By active present and future tense verbs, the person and number can be recognized thanks to the inflectional suffix (excluding *-í* suffix which can be used to indicate both 3rd person singular and plural). By active past tense verbs or passive verbs, the gender can be also specified (excluding *-a* suffix indicating either active past tense 3rd feminine singular or neutrum plural).

Jane likes to bake. Today she has baked an apple-pie.

Figure 1. Zero pronominal anaphora: the implicit subject *Ø* refers to *Jane*.

We used the maximum entropy method to train a model for unstated subject classification and chose the data of the PDT 2.0 for the training and testing procedures. However, the corpus selection does not suit the task and we will discuss it later.

Motivation

In machine translation, the identification and resolution of zero pronouns play an important role if these pronouns are often omitted in the source language (e.g. Czech) but compulsory in the target language (e.g. English). There are systems for anaphora resolution in Czech [*e.g., Kučová & Žabokrtský,* 2005; *Nguy & Žabokrtský,* 2007; *Nguy et al.,* 2009], but none of them devoted sufficient space to zero pronoun resolution in particular as well as zero pronoun identification. Nevertheless, zero pronoun detection is an unseparatable part of anaphora resolution and has been widely investigated in other languages, e.g. in Japanese [*Seki et al.,* 2002], Chinese [*Zhao & Ng,* 2007], Korean [*Han et al.,* 2006], Spanish [*Ferrández & Peral,* 2000] etc.

Unstated Subject Identification

In this section, we introduce the method and the corpus we have used for unstated subject identification. We discuss the problems we have met during our experiments.

Resolution method

Maximum entropy was first introduced to Natural Language Processing (NLP) area by *Berger et al.* [1996]. Since then, the maximum entropy principle has been used widely in NLP, e.g. for tagging, parsing, named entity recognition and machine translation. Maximum entropy models have the following form:

$$p(y|x) = \frac{1}{Z(x)} exp \sum_i \lambda_i f_i(x, y)$$

where f_i is a feature function, λ_i is its weight, and $Z(x)$ is the normalizing factor.

For our task, we chose a maximum entropy classifier, an implementation of Laye Suen[3], a machine learning tool that takes data items and place them into one of *k* classes. In addition, it also gives probability distributions over classifications.

Our approach can be described in the following steps:

1. In a training set, extract features from each finite verb without an overt subject;
2. Train a MaxEnt classifier with them;
3. Test the MaxEnt model on a test set;

Data description

Our experiments make use of the PDT 2.0, which contains a large amount of Czech newspaper texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level [*Hajič et al.,* 2006].

At the so-called tectogrammatical (semantic) layer (t-layer), the meaning of the sentence is represented as a dependency tree structure. In addition to nodes corresponding to surface tokens, there are newly established nodes the tectogrammatical lemma of which is an artificial t-lemma substitute beginning with *#*. Our focused unstated subjects can be found at t-layer among nodes with t-lemma `#PersPron`, `#Gen` and `#Unsp`; except null subjects, which were not reconstructed at t-layer. These t-lemma substitutes have the following meanings:

`#PersPron` t-lemma substitutes are assigned to:

- personal and possessive pronouns present in the surface sentence;
- zero pronouns representing the implicit subject[4];
- textual ellipsis – obligatory arguments of a governing verb / noun[5];

[3] A Perl module `AI::MaxEntropy`, see http://search.cpan.org/perldoc?AI::MaxEntropy

[4] Dropped pronouns can be distinguished from the expressed ones by the additional node attribute is_generated

[5] A node with the t-lemma substitute `#PersPron` is used in cases of textual ellipsis no matter what the form of the omitted argument is; i.e. not only in the positions where it could be replaced by a personal or possessive pronoun.

#Gen t-lemma substitutes are used for:

- grammatical ellipsis of an obligatory argument – general argument;
- zero pronouns representing the general subject;

#Unsp t-lemma substitutes stand for:

- grammatical ellipsis of an obligatory argument – unspecified Actor;
- zero pronouns representing the unspecified subject;

Feature extraction

Our maximum entropy classifier was trained on the basis of feature vectors for each finite verb (predicate) having no overt subject depending on it. The following features were used:

Categorial features: t-lemma, form, tense, gender, number, person, and:

- adverbial form – an adverb in the case of an 'adverbial' predicate ('to be + an adverb')
- nominal form – a nominal part in the case of a nominal predicate

Binary features:

- `has_actor` – the considered predicate has an overt Actor
- `is_reflexive` – the predicate is reflexive
- `is_passive` – the predicate is a passive verb
- `has_o-ending` – the predicate is a finite verb ending with 'o'
- `is_to-be-infin` – the predicate is in the construction of 'to be + infinitive'
- `has_dep-clause` – there is a dependent clause hanging on the verb

Concatenated features:

- `reflexive_o-ending` – concatenation of the features is_reflexive and has_o-ending
- `passive_o-ending` – concatenation of the features is_passive and has_o-ending
- `reflexive_person_number_gender` – concatenation of the features is_reflexive, person, number and gender
- `passive_person_number_gender` – concatenation of the features is_passive, person, number and gender

The feature selection relies on characteristics of each unstated subject type. A general subject often comes along with a third person singular reflexive verb or a third person singular passive verb. A reflexive verb can be easily recognized by a reflexive particle. A third person singular passive verb and a past tense third person singular reflexive verb always end with 'o'. The case of a subject expressed by a dependent clause can be detected by the has_dep-clause feature. An adverbial form can indicate a null subject, e.g. *Je polojasno* ('It is somewhat cloudy').

Data problems

By the PDT 2.0 choice we have to face the problems related to it. The most crucial problem is the absence of the explicit annotation of unstated subjects we interest in. In Fig. 2 and Fig. 3, we illustrate ambiguous cases, in which two nodes with #PersPron and #Gen appear.

We tried to solve the problem of missing manual unstated subject annotation by proposing some rules:

if [the verb has a #Unsp among children] **then**
    It is the case of an unspecified subject
else if [the verb has a generated #PersPron and a #Gen.ACT among children] **then**
    if [the verb has_o-ending **or** is_to-be-infin **or** is_rflx_pass_by_active_present_3sg] **then**
        It is the case of a general subject
    else
        It is the case of an implicit subject
    end if
else if [the verb has a generated #PersPron.ACT among children] **then**
    It is the case of an implicit subject
else if [the verb has a #Gen.ACT among children] **then**
    It is the case of a general subject
else if [[the verb has a generated #PersPron.ACT among children] **and**
    [[it is_pass **or** is_rflx_pass_not_active_present_3sg] with no o-ending]] **then**
    It is the case of an implicit subject
else
    It is the case of a null subject

Another problem with the PDT 2.0 data is the absence of the manual annotation of person, number and gender. This information is very important for us because of its indication for a general / null subject by a third

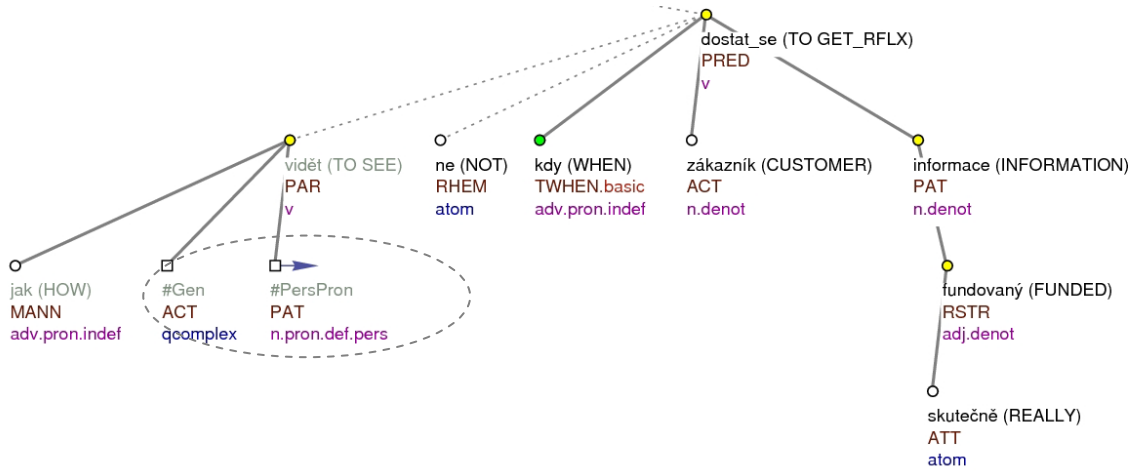


Figure 2. A simplified t-tree representing the sentence *Jak je vidět, ne vždy se zákazníkovi dostane skutečně fundovaných informací.* (Lit. How it's seen, not always RFLX customer gets really funded information.) In this case, the node with `#Gen` is considered to be the unstated general subject.

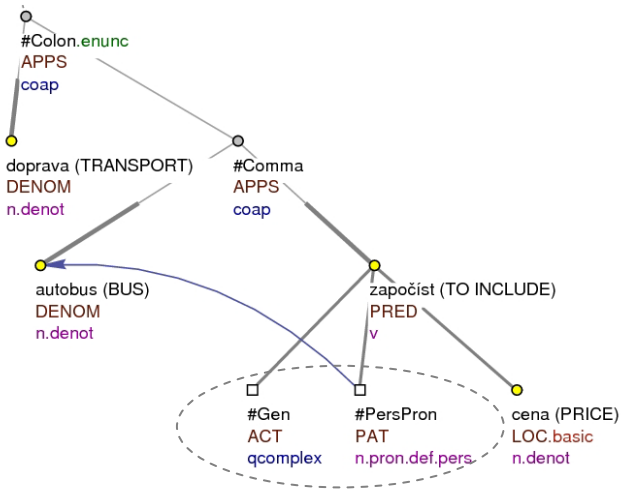


Figure 3. A simplified t-tree representing the sentence *Doprava: Autobus, je započten v ceně.* (Lit. Transport: Bus, is included in price.) In this case, the node with `#PersPron` is the unstated implicit subject.

person singular neuter / animate form or for an unspecified subject by a third person plural animate form.

We have no rules that guarantee a 100% correct resolution for the identification of unstated subjects on annotated data of the PDT 2.0. In addition, we rely on the genre of the corpus, where proverbs with general subjects do not often occur, and suppose all cases with third person singular animate active verb to be an implicit subject; whereas all cases with third person singular neutrum passive or reflexive verb to be a general subject. We expect that the occurrence of singular neuter implicit subject is sporadic as well.

Baselines

The following baselines were created for automatic identification of:

Implicit subject:

if [[a clause contains a finite verb] **and**
[the verb has neither overt subject nor Actor depending on it] **and**
[it has no o-ending] **and** [it is not a reflexive passive verb] **and**
[it is not a passive verb with o-ending] **and** [its t-lemma is not an impersonal verb[6]]
] **then**
There is an implicit subject.
end if

[6] We have manually created a test list of impersonal verbs consisting of verbs: *jednat se* (be about sth), *pršet* (rain), *zdát se* (seem), *dařit se* (do well), *oteplovat se* (get warmer), *ochladit se* (get colder)

General subject:
if [[a clause contains a finite verb] **and**
[the verb has neither overt subject nor Actor depending on it] **and**
[[it has o-ending] **or** [it has a 'to be + infinitive' construction] **or**
[it is a reflexive passive verb having an active present tense third person singular form]
]
] **then**
There is a general subject.
end if
Unspecified subject:
if [[a clause contains a finite verb] **and**
[the verb has no overt subject depending on it and a third person animate plural form] **and**
[[there is no preceding finite verb] **or**
[[there is a preceding finite verb] **and**
[[it has not a third person animate plural form] **or**
[it has not a dependent animate plural noun with functor ACT/PAT/ADDR]
]
]
]
] **then**
There is an unspecified subject.
end if
Null subject:
if [[a clause contains a finite verb] **and**
[the verb has neither overt subject nor Actor depending on it] **then**
There is a null subject.
end if

Evaluation

The PDT 2.0 is divided into three parts: 80% of data is used for training, 10% for development testing and 10% for evaluation testing. In the evaluation, we used the standard metrics with precision, recall and f-measure (Table 1) for unstated subject identification.

Precision = N_c / N_e	Recall = N_c / N_t	F-measure = 2 x P x R / (P + R)

N_c	Number of correctly identified controllees
N_e	Number of identified controllees
N_t	Number of all controllees

Table 1. Evaluation metrics for the unstated subject identification

If the problem of missing manual unstated subject annotation is considered to be 100% successfully resolved by proposed hand-written rules, then we obtain the following results (Table 2):

	P	R	F
Implicit Baseline	95.4%	98.4%	96.9%
Implicit MaxEnt	90.6%	99.4%	94.8%
General Baseline	24.9%	87.2%	38.7%
General MaxEnt	96.7%	74.4%	84.1%
Unspecified Baseline	4.55%	3.45%	3.92%
Unspecified MaxEnt	0%	0%	0%
Null Baseline	98%	85.7%	91.5%
Null MaxEnt	82.5%	29.7%	43.7%

Table 2. Results for the unstated subject identification

Such a poor result of unspecified subject identification can be explained for its rare occurrences in the data, the problem of missing manual person, gender and number annotation and the fact that it requires knowledge of a potential antecedent existence. If there is an antecedent to which the unstated subject can refer, then it is a case of an implicit subject; otherwise an unspecified subject. An anaphora resolution might help to improve this result.

The result of null subject identification might be higher by adding a sophisticated list of impersonal / weather verbs / constructions as well. In general a deeper error analysis should bring overall improvements and explain the doubt of better baseline results.

Conclusion

This paper introduces the linguistics phenomenon of unstated subjects in Czech and its automatic identification using a maximum entropy classifier trained on the PDT 2.0 data. Looking at the results and the data problems leads us to a question, whether we should continue on improving the approach to unstated subject identification or whether we should concern on the automatic identification of `#PersPron`, `#Gen` and `#Unsp` nodes as specified in the PDT 2.0 instead. Or should we try to find another more suitable corpus for the task?

Acknowledgments. The present work was supported by the Czech Grant Agency under Contracts GA CR P406/2010/0875, MSMT CR LC536 and by the Charles University Grant Agency under Contract GAUK 4383/2009.

References

Berge, A. L., V. J. Della Pietra, and S. A. Della Pietra, A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, 22(1), 39–71, 1996.

Ferrández, A. and J. Peral. A Computational Approach to Zero-pronouns in Spanish, in: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, 166–172, Morristown, NJ, USA, Association for Computational Linguistics, 2000.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský and M. Ševčíková-Razímová, *Prague Dependency Treebank 2.0*, Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, www.ldc.upenn.edu, 2006.

Han N. R., E. F. Prince, M. Palmer, and E. Buckley, *Korean Zero Pronouns: Analysis and Resolution*. Ph.D. thesis, University of Pennsylvania, 2006.

Kučová, L. and Z. Žabokrtský, Anaphora in Czech: Large Data and Experiments with Automatic Anaphora, *LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue*, 3658:93–98, 2005.

Nguy, G. L. and Z. Žabokrtský, Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data, in: *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, 77–81, Lagos (Algarve), Portugal, CLUP-Center for Linguistics of the University of Oporto, 2007.

Nguy, G. L., V. Novák and Z. Žabokrtský, Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech, in: *Proceedings of the SIGDIAL 2009 Conference*, 276–285, The Association for Computational Linguistics, London, UK, 2009.

Rello, L. and I. Ilisei, A Rule-Based Approach to the Identification of Spanish Zero Pronouns, in: *Proceedings of Student Workshop of Recent Advances in Natural Language Processing (RANLP 2009)*, 60–65, Borovets, Bulgaria, 14-15 September, 2009.

Seki, K., A. Fujii, and T. Ishikawa, A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution, in: *Proceedings of the 19th International Conference on Computational Linguistics*, 1–7, Morristown, NJ, USA, Association for Computational Linguistics, 2002.

Zhao, S. and H. T. Ng, Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach, in: *Proceedings of the 2007Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.