Learning Information Structure in The Prague Treebank

Oana Postolache
Department of Computational Linguistics
University of Saarland, Saarbrücken, Germany
oana@coli.uni-sb.de

Abstract

This paper investigates the automatic identification of aspects of Information Structure (IS) in texts. The experiments use the Prague Dependency Treebank which is annotated with IS following the Praguian approach of Topic Focus Articulation. We automatically detect t(opic) and f(ocus), using node attributes from the treebank as basic features and derived features inspired by the annotation guidelines. We show the performance of C4.5, Bagging, and Ripper classifiers on several classes of instances such as nouns and pronouns, only nouns, only pronouns. A baseline system assigning always f(ocus) has an F-score of 42.5%. Our best system obtains 82.04%.

1 Introduction

Information Structure (IS) is a partitioning of the content of a sentence according to its relation to the discourse context. There are numerous theoretical approaches describing IS and its semantics (Halliday, 1967; Sgall, 1967; Vallduví, 1990; Steedman, 2000) and the terminology used is diverse (see (Kruijff-Korbayová & Steedman, 2003) for an overview). However, all theories consider at least one of the following two distinctions: (i) a topic/focus[1] distinction that divides the linguistic meaning of the sentence into parts that link the content to the context, and others that advance the discourse, i.e. add or modify information; and (ii) a background/kontrast[2] distinction between parts of the utterance which contribute to distinguishing its actual content from alternatives the context makes available. Existing theories, however, state their principles using carefully selected illustrative examples. Because of this, they fail to adequately explain what possibly different linguistic dimensions cooperate to realize IS and how they do it.

In this paper we report the results of an experiment aimed to automatically identify aspects of IS. This effort is part of a larger investigation aimed to get a more realistic view on the realization of IS in naturally occurring texts.

For such an investigation, the existence of a corpus annotated with some kind of 'informativity status' is of great importance. Fully manual annotation of such a corpus is tedious and time-consuming. Our plan is to initially annotate a small amount of data and then to build models to automatically detect IS in order to apply bootstrapping techniques to create a larger corpus.

This paper describes the results of a pilot study; its aim is to check if the idea of learning IS works by trying it on an already existing corpus. For our experiments, we have used the Prague Dependency Treebank (PDT) (Hajič, 1998), as it is the only corpus annotated with IS (following the theory of Topic-Focus Articulation). We trained three different classifiers, C4.5, Bagging and Ripper, using basic features from the treebank and derived features inspired by the annotation guidelines. We have evaluated the performance of the classifiers against a baseline that simulates the preprocessing procedure that preceded the manual annotation of PDT, and

[1] We use the Praguian terminology for this distinction.

[2] The notion 'kontrast' with a 'k' has been introduced in (Vallduví and Vilkuna, 1998) to replace what Steedman calls 'focus', and to avoid confusion with other definitions of focus.

Proceedings of the ACL Student Research Workshop, pages 115–120,
Ann Arbor, Michigan, June 2005.

against a rule-based system which we implemented following the annotation instructions.

The organization of the paper is as follows. Section 2 describes the Prague Dependency Treebank, Section 3 presents the Praguian approach of Topic-Focus Articulation, from two perspectives: of the theoretical definition and of the annotation guidelines that have been followed to annotate the PDT. Section 4 presents the experimental setting, evaluation metric and results. The paper closes with conclusions and issues for future research (Section 5).

2 Prague Dependency Treebank

The Prague Dependency Treebank (PDT) consists of newspaper articles from the Czech National Corpus (Čermaák, 1997) and includes three layers of annotation. **The morphological layer** gives a full morphemic analysis in which 13 categories are marked for all sentence tokens (including punctuation marks). **The analytical layer**, on which the "surface" syntax (Hajič, 1998) is annotated, contains analytical tree structures, in which every token from the surface shape of the sentence has a corresponding node labeled with main syntactic functions like SUBJ, PRED, OBJ, ADV. **The tectogrammatical layer** renders the deep (underlying) structure of the sentence (Sgall et al., 1986; Hajičová et al., 1998). Tectogrammatical tree structures (TGTSs) contain nodes corresponding only to the autosemantic words of the sentence (e.g., no preposition nodes) and to deletions on the surface level; the condition of projectivity is obeyed, i.e. no crossing edges are allowed; each node of the tree is assigned a functor such as ACTOR, PATIENT, ADDRESSEE, ORIGIN, EFFECT, the list of which is very rich; elementary coreference links are indicated, in the case of pronouns.

3 Topic Focus Articulation (TFA)

The tectogrammatical level of the PDT was motivated by the more and more obvious need of large corpora that treat not only the morphological and syntactic structure of the sentence but also semantic and discourse-related phenomena. Thus, TGTSs have been enriched with features displaying the information structure of the sentence which is a means of showing its contextual potential.

3.1 Theory

In the Praguian approach to IS, the content of the sentence is divided in two parts: the Topic is "what the sentence is about" and the Focus represents the information asserted about the Topic. A prototypical declarative sentence asserts that its Focus holds (or does not hold) about its Topic: Focus(Topic) or not-Focus(Topic).

The TFA definition uses the distinction between Context-Bound (CB) and Non-Bound (NB) parts of the sentence. To distinguish which items are CB and which are NB, the question test is applied, (i.e., the question for which a given sentence is the appropriate answer is considered). In this framework, weak and zero pronouns and those items in the answer which reproduce expressions present (or associated to those present) in the question are CB. Other items are NB.

In example (1), (b) is the sentence under investigation, in which CB and NB items are marked, (a) is the context in which the sentence is uttered, and (c) is the question for which the given sentence is an appropriate answer:

(1) (a) Tom and Mary both came to John's party.
(b) John$_{CB}$ invited$_{CB}$ only$_{NB}$ her$_{NB}$.
(c) Whom did John invite?

The following rules determine which lexical items (CB or NB) belong to the Topic or to the Focus (Hajičová et al., 1998; Hajičová and Sgall, 2001):

1. The main verb and any of its direct dependents belong to the Focus if they are NB;
2. Every item that does not depend directly on the main verb and is subordinated to an element of Focus belongs to Focus (where "subordinated to" is defined as the irreflexive transitive closure of "depend on");
3. If the main verb and all its dependents are CB, then those dependents k_i of the verb which have subordinated items l_m that are NB are called 'proxi foci'; the items l_m together with all items subordinated to them belong to Focus, where $i, m > 1$;
4. Every item not belonging to Focus according to 1 – 3 belongs to Topic.

3.2 Annotation guidelines

Within PDT, the TFA attribute has been annotated for all nodes (including the restored ones) at the tectogrammatical level. Instructions for the assignment of TFA attribute have been specified in (Buráňová et al., 2000) and are summarized in Table 1. These instructions are based on the surface word order, the position of the sentence stress (intonation center – IC[3]) and the canonical order of the dependents.

The TFA attribute has 3 values: t, for non-contrastive CB items; f, for NB items; and c, for contrastive CB items. In this paper, we do not distinguish between contrastive and non-contrastive items, considering both of them as being just t. In the PDT annotation, the values t (from topic) and f (from focus) have been chosen to be used because, in the most cases, in prototypical sentences, t items belong to the Topic and f items to the Focus.

Before the manual annotation, the corpus has been preprocessed to mark all nodes with the TFA attribute of f, as it is the more common value. Then the annotators changed the value according to the guidelines in Table 1.

4 Automatic extraction of TFA

In this section we consider the automatic identification of t and f using machine learning techniques trained on the annotated data.

The data set consists of 1053 files (970,920 words) from the pre-released version of PDT 2.0.[4] We restrict our experiments by considering only noun- and pronoun-nodes. The total number of instances (nouns and pronouns) in the data is 297,220 out of which 254,242 (86.54%) are nouns and 39,978 (13.46%) are pronouns. The t/f distribution of these instances is 172,523 f (58.05%) and 124,697 t (41.95%).

We experimented with three different classifiers, C4.5, Bagging and Ripper, because they are based on different machine learning techniques (decision trees, bagging, rules induction) and we wanted to see which of them performs better on this task. We used Weka implementations of these classifiers (Witten and Frank, 2000).

4.1 Features

The experiments use two types of features: (1) basic features of the nodes taken directly from the treebank (node attributes), and (2) derived features inspired by the annotation guidelines.

The basic features are the following (the first 4 are boolean, and 5 and 6 are nominal):

1. **is-noun**: true, if the node is a noun;
2. **is-root**: true, if the node is the root of the tree;
3. **is-coref-pronoun**: true, if the node is a coreferential pronoun;
4. **is-noncoref-pronoun**: true, if the node is a non-coreferential pronoun (in Czech, many pronouns are used in idiomatic expressions in which they do not have an coreferential function, e.g., *svého času*, lit. 'in its (reflexive) time', 'some time ago');
5. **SUBPOS**: detailed part of speech which differentiates between types of pronouns: personal, demonstrative, relative, etc.;
6. **functor**: type of dependency relations: MOD, MANN, ATT, OTHER.

The derived features are computed using the dependency information from the tectogrammatical level of the treebank and the surface order of the words corresponding to the nodes[5]. Also, we have used lists of forms of Czech pronouns that are used as weak pronouns, indexical expressions, pronouns with general meaning, or strong pronouns. All the derived features have boolean values:

7. **is-rightmost-dependent-of-the-verb**;
8. **is-rightside-dependent-of-the-verb**;
9. **is-leftside-dependent**;
10. **is-embedded-attribute**: true, if the node's parent is not the root;
11. **has-repeated-lemma**: true, in case of nouns, when another node with the same lemma appears in the previous 10 sentences.
12. **is-in-canonical-order**;
13. **is-weak-pronoun**;
14. **is-indexical-expression**;
15. **is-pronoun-with-general-meaning**;
16. **is-strong-pronoun-with-no-prep**;

[3] In the PDT the intonation center is not annotated. However, the annotators were instructed to use their opinion where the IC is when they utter the sentence.

[4] We are grateful to our colleagues at the Charles University in Prague for providing us the experimental data before the PDT 2.0 official release.

[5] On the tectogramatical level in the PDT, the order of the nodes has been changed during the annotation process of the TFA attribute, so that all t items precede all f items. Our features use the surface order of the words corresponding to the nodes.

1.	The bearer of the IC (typically, the rightmost child of the verb)	f
2.	If IC is not on the rightmost child, everything after IC	t
3.	A left-side child of the verb (unless it carries IC)	t
4.	The verb and the right children of the verb before the f-node (cf. 1) that are canonically ordered	f
5.	Embedded attributes (unless repeated or restored)	f
6.	Restored nodes	t
7.	Indexical expressions (*já* I, *ty* you, *těd* now, *tady* here), weak pronouns, pronominal expressions with a general meaning (*někdo* somebody, *jednou* once) (unless they carry IC)	t
8.	Strong forms of pronouns not preceded by preposition (unless they carry IC)	t

Table 1: Annotation guidelines; IC = Intonation Center

4.2 Evaluation framework

In order to perform the evaluation, we randomly selected 101,054 instances (1/3 of the data) from all the instances, which represents our test set; the remaining 2/3 of the data we used as a training set. The same test set is used by all three classifiers. In our experiments we have not tweaked the features and thus we have not set aside a development set. In the test set 87% of the instances are nouns and 13% are pronouns. The t/f distribution in the test set is as follows: 58% of the instances are t, and 42% instances are f.

We have built models using decision trees (C4.5), bagging and rule-induction (Ripper) machine learning techniques to predict the Information Structure.

We have also implemented a deterministic, rule-based system that assigns t or f according to the annotation guidelines presented in Table 1. The rule-based system does not have access to what intonation center (IC) is.

The baseline simulates the preprocessing procedure used before the manual annotation of TFA attribute in the PDT, i.e., assigns always the class that has the most instances.

Our machine learning models are compared against the baseline and the rule-based system. As a metric we have used the Weighted Averaged F-score which is computed as follows:

```
%_f*F-score_f+%_t*F-score_t
```

The reason why we have chosen this metric (instead of Correctly Classified, for example) is that it gives a more realistic evaluation of the system, considering also the distribution of t and f items [6].

[6] Consider, for example, the case in which the test set consists of 70% f items and 30% t items. The Baseline system would have as much as 70% correctly classified instances, just because the t/f distribution is as such. The Weighted Averaged F-score would be in this case 57.64% which is a more adequate value that reflects better the poorness of such a system.

4.3 Results

The results of the experiment using all instances (nouns and pronouns) are shown in Table 2 in the second column. C4.5 and Bagging achieve the best performance improving on the results of the rule-based system by 6.99%.

The top of the decision tree generated by C4.5 in the training phase looks like this:

```
is-coref-pronoun = true
|    is-leftside-dependent = true
|    |    SUBPOS = ...
is-coref-pronoun = false
|    is-leftside-dependent = true
|    |    is-in-canonical-order = true
```

The overall tree has 129 leaves out of 161 nodes.

In order to achieve a better understanding of the difficulty of the task for nouns and pronouns, we considered evaluations on the following classes of instances:

- only nouns;
- nouns that are direct dependents of the verb (verb children);
- nouns that are not direct dependents of the verb (non-verb children);
- only pronouns;
- coreferential pronouns;
- non-coreferential pronouns.

We also wanted to investigate if the three classifiers perform differently with respect to different classes of instances (in which case we could have a general system, that uses more classifiers, and for certain classes of instances we would 'trust' a certain classifier, according to its performance on the development data).

Systems	nouns & pronouns	only nouns			only pronouns		
		all	verb children	non-verb children	all	coref	non-coref
Baseline	42.50	51.43	41.90	73.08	81.35	96.94	58.79
Rule-based	76.68	75.59	79.09	69.06	82.23	95.51	62.44
C4.5	82.04	79.98	80.38	73.87	93.77	97.25	68.60
Bagging	82.04	79.97	80.37	73.86	93.71	97.34	68.36
Ripper	81.78	79.88	80.31	73.86	93.55	97.35	68.36

Table 2: Overall results: Weighted Averaged F-score as percentage

Table 2, in columns three and onwards, shows the results on different classes of instances. The test set for each class of instances represents 1/3 randomly extracted instances from all instances in the data belonging to that class, in the same fashion as for the overall split.

The baseline is higher for some classes, yet the classifiers perform always better, even than the rule-based system, which for non-verb children performs worse than the baseline. However, the difference between the three classifiers is very small, and only in one case (for the coreferential pronouns) C4.5 is outperformed by Ripper.

To improve the results even more, there are two possibilities: either providing more training data, or considering more features. To investigate the effect of the size of the training data we have computed the learning curves for the three classifiers. Figure 1 shows the C4.5 learning curve for the overall experiment on nouns and pronouns; the learning curves for the other two classifiers are similar, and not included in the figure.

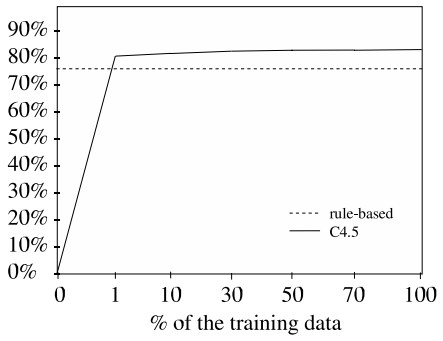

Figure 1: Learning curve for the C4.5 classifier

The curve is interesting, showing that after only 1% of the training set (1961 instances) C4.5 can already perform well, and adding more training data improves the F-score only slightly. To ensure the initial 1% aren't over-representative of the kind of IS phenomena, we experimented with different 1% parts of the training set, and the results were similar. We also did a 10-fold cross validation experiment on the training set, which resulted in a Weighted Averaged F-score of 82.12% for C4.5.

The slight improvement achieved by providing more data indicates that improvements are likely to come from using more features.

Table 3 shows the contribution of the two types of features (basic and derived) for the experiment with all instances (nouns and pronouns). For comparison we have displayed again the baseline and the rule-based system F-score.

Features / System	Basic	Derived	All
C4.5	62.82	77.51	82.04
Bagging	62.83	77.50	81.99
Ripper	62.48	77.28	81.78
Rule-based	76.68		
Baseline	42.50		

Table 3: Contribution of different features. F-score given as a percentage.

The results show that the model trained only with basic features performs much better than the baseline, yet it is not as good as the rule-based system. However, removing the basic features completely and keeping only the derived features considerably lowers the score (by more than 4%). This indicates that adding more basic features (which are easy to obtain from the treebank) could actually improve the results.

The derived features, however, have the biggest impact on the performance of the classifiers. Yet, adding more sophisticated features that would help in this task (e.g., coreferentiality for nouns) is difficult because they cannot be computed reliably.

5 Conclusions

In this paper we investigated the problem of learning aspects of Information Structure from annotated data. We presented results from a study trying to verify whether Information Structure can be learned using mostly syntactic features. We used the Prague Dependency Treebank which is annotated with IS following the Praguian theory of Topic Focus Articulation. The results show that we can reliably identify t(opic) and f(ocus) with over 82% Weighted Averaged F-score while the baseline is at 42%.

Issues for further research include, on the one hand, a deeper investigation of the Topic-Focus Articulation in the Prague Dependency Treebank, by improving the feature set, considering also the distinction between contrastive and non-contrastive t items and, most importantly, by investigating how we can use the t/f annotation in PDT (and respectively our results) in order to detect the Topic/Focus partitioning of the whole sentence.

On the other hand, we want to benefit from the experience with the Czech data in order to create an English corpus annotated with Information Structure. We want to exploit a parallel English-Czech corpus available as part of the PDT, in order to extract correlations between different linguistic dimensions and Topic/Focus in the Czech data and investigate how they can be transferred to the English version of the corpus.

References

Eva Buránová, Eva Hajičová & Petr Sgall. 2000. *Tagging of very large corpora: Topic-Focus Articulation.* Proceedings of the 18th International Conference on Computational Linguistics, COLING 2000, 139-144.

František Čermák. 1997. *Czech National Corpus: A Case in Many Contexts.* International Journal of Corpus Linguistics, 2(2):181-197.

Jan Hajič. 1998. *Building a syntactically annotated corpus: The Prague Dependency Treebank.* Issues of valency and Meaning. Studies in Honor of Jarmila Panevová, ed. by E. Hajičová. Karolinum, Prague.

Eva Hajičová, Barbara Partee & Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content.* Studies in Linguistics and Philosiphy, 71 Dordrecht: Kluwer.

Eva Hajičová & Petr Sgall. 2001. *Topic-focus and salience.* Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL 2001, 268-273. Toulose, France.

M. Halliday. 1967. *Notes on transitivity and theme in English, Part II.* Journal of Linguistic, 3:199-244.

Ivana Kruijff-Korbayová and Mark Steedman. 2003. *Discourse and Information Structure.* Journal of Logic, Language and Information 12:249-259. Kluwer, Amsterdam.

Petr Sgall. 1967. *Functional sentence perspective in a generative description.* Prague Studies in Mathematical Linguistics, 2:203-225.

Petr Sgall, Eva Hajičová & Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Reidel, Dordrecht.

Mark Steedman. 2000. *Information Structure and the syntax-phonology interface.* Linguistic Inquiry, 34:649-689.

Enrich Vallduví. 1990. *The information component.* Ph.D Thesis, University of Pennsylvania.

Enric Vallduví & Maria Vilkuna. 1998. *On rheme and kontrast.* Syntax and Semantics, Vol. 29: The Limits of Syntax, ed. by P. Culicover and L. McNally. Academic Press, San Diego.

Ian H. Witten & Eibe Frank. 2000. *Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.