Automatically Acquiring Fine-Grained Information Status Distinctions in German

Aoife Cahill
Educational Testing Service,
660 Rosedale Road,
Princeton, NJ 08541, USA
acahill@ets.org

Arndt Riester
Institute for Natural Language Processing (IMS)
Pfaffenwaldring 5b
70569 Stuttgart, Germany
arndt.riester@ims.uni-stuttgart.de

Abstract

We present a model for automatically predicting information status labels for German referring expressions. We train a CRF on manually annotated phrases, and predict a fine-grained set of labels. We achieve an accuracy score of 69.56% on our most detailed label set, 76.62% when gold standard coreference is available.

1 Introduction

The automatic identification of *information status* (Prince, 1981; 1992), i.e. categorizing discourse entities into different classes on the *given-new* scale, has recently been identified as an important issue in natural language processing (Nissim, 2006; Rahman and Ng, 2011; 2012). It is widely acknowledged that information status and, more generally, information structure,[1] is reflected in word order, in the form of referring expressions as well as in prosody. In computational linguistics, the ability to automatically label text with information status, therefore, could be of great benefit to many applications, including surface realization, text-to-speech synthesis, anaphora resolution, summarization, etc.

The task of automatically labeling text with information status, however, is a difficult one. Part of the difficulty arises from the fact that, to a certain degree, such labeling requires world knowledge and semantic comprehension of the text, but another obstacle is simply that theoretical notions of information status are not used consistently in the literature.

In this paper we outline a system, trained on a small amount of data, that achieves encouraging results on the task of automatically labeling transcribed German radio news data with fine-grained information status labels.

2 Learning information status

A simpler variant of the task is *anaphoricity detection* (discourse-new detection) (Bean and Riloff, 1999; Ng and Cardie, 2002; Uryupina, 2003; Denis and Baldridge, 2007; Zhou and Kong, 2011), which divides discourse entities into *anaphoric (given)* and *new*. Identifying discourse-new expressions in texts is helpful as a precursor to coreference resolution, since, by definition, there is no need to identify antecedents for new entities.

In the linguistic literature, referring expressions have been distinguished in much more detail, and there is reason to believe that this could also provide useful information for NLP applications. Nissim (2006) and Rahman and Ng (2011) developed methods to automatically identify three different classes: OLD, MEDIATED and NEW expressions. This classification, which is described in Nissim et al. (2004), has been used for annotating the *Switchboard* dialog corpus (Calhoun et al., 2010), on which both studies are based. Most recently, Rahman and Ng (2012) extend their automatic prediction system to a more fine-grained set of 16 subtypes.

[1] Information structure is usually taken to describe clause-internal divisions into *focus-background, topic-comment, or theme-rheme*, which are in turn defined in terms of contextual factors such as *given-new* information, *salience*, *contrast* and *alternatives*, cf. Steedman and Kruijff-Korbayová (2003), Krifka (2007). *Information status* is the subfield of information structure which exclusively deals with the *given-new* distinction and which is normally confined to referring expressions.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 232–236,
Seoul, South Korea, 5-6 July 2012.

Old. The class of OLD entities in Nissim et al. (2004) is not limited to full-fledged anaphors like in Example (1a) but also includes cases of generic and first/second person pronouns like in (1b), which may or may not possess a previous mention.

(1) a. Shares in *General Electric* rose as investors bet that the US company would take more lucrative engine orders for the A380.
b. I wonder where this comes from.

Mediated. The group of MEDIATED entities mainly has two subtypes: (2a) shows an expression which has not been mentioned before but which is dependent on previous context. Such items have also been called *bridging anaphors* (Poesio and Vieira, 1998). (2b) contains a phrase which is generally known but does not depend on the discourse context.

(2) a. Tomorrow, *the Shenzhou 8 spacecraft* will be in a position to attempt the docking.
b. They hope that he will be given the right to remain in the Netherlands.

New. The label NEW, following Nissim et al. (2004: 1024), applies "to entities that have not yet been introduced in the dialog and that the hearer cannot infer from previously mentioned entities."[2] Two kinds of expressions which fall into this category are unfamiliar definites (3a) and (specific) indefinites (3b).

(3) a. The man who shot a policeman yesterday has not been caught yet.
b. Klose scored a penalty in the 80th minute.

Based on work described in Nissim (2006), Rahman and Ng (2011) develop a machine learning approach to information-status determination. They develop a support vector machine (SVM) model from the annotated Switchboard dialogs in order to predict the three possible classes. In an extension of this work, Rahman and Ng (2012) compare a rule-based system to a classifier with features based on the rules to predict 16 subtypes of the three basic types. On this extended label set on the dialog data, they achieve accuracy of 86.4% with gold standard coreference and 78.7% with automatically detected coreference.

3 Extending Information Status prediction

The work we present here is most similar to that of Rahman and Ng (2012), however, our work differs from theirs in a number of important respects. We (i) experiment with a different information status classification, derived from Riester et al. (2010), (ii) use (morpho-)syntactic and functional features automatically extracted from a deep linguistic parser in our CRF sequence model, (iii) test our approach on a different language (German), (iv) show that high accuracy can be achieved with a limited number of training examples, and (v) that the approach works on a different genre (transcribed radio news bulletins which contain complex embedded phrases like *an offer to the minority Tamil population of Sri Lanka*, not typically found in spoken dialog).

The annotation scheme by Riester et al. (2010) divides referring items differently to Nissim et al. (2004). Arguments are provided in the former paper and in Baumann and Riester (to appear). As it stands, the scheme provides too many labels for our purpose. As a compromise, we group them in seven classes: GIVEN, SITUATIVE, BRIDGING, UNUSED, NEW, GENERIC and EXPLETIVE.

Given. *Givenness* is a central notion in information structure theory. Schwarzschild (1999) defines givenness of individual-type entities in terms of coreference. If desired, GIVEN items can be subclassified, e.g. whether they are pronouns or full noun phrases, and whether the latter are repetitions or short forms of earlier material, or whether they consist of lexically new material (epithets).

Situative. 1^{st} and 2^{nd} person pronouns, locative and temporal adverbials, usually count as deictic expressions since they refer to elements in the utterance situation. We therefore count them as a separate class. SITUATIVE entities may, but need not, corefer.

Bridging. Bridging anaphors, as in (2a) above, have received much attention, see e.g. Asher and Lascarides (1998) or Poesio and Vieira (1998). Although they are discourse-new, they share properties with coreference anaphors since they depend on the discourse context. They represent a class which can be easily identified by human annotators but are difficult to capture by automatic techniques.

Unused. In manual annotation practice, it is very often impossible to decide whether an entity is hearer-known, since this depends on who we assume the hearer to be; and even if we agree on a recipient, we may still be mistaken about their knowledge. For example, *Wolfgang Bosbach, deputy chairman of the*

[2] Note that this definition fails to exclude cases like (2b).

Countable	Boolean	Descriptive
# Words in phrase*	Phrase contains a compound noun	Adverbial type, e.g. locative
# Predicative phrases	Phrase contains coordination	Determiner type, e.g. definite *
# DPs and NPs in phrase	Phrase contains time expression	Left/Right-most POS tag of phrase
# top category children	Phrase contains $<$ *2, 5* or *10* words	Highest syntactic node label
# Labels/titles	Phrase does not have a complete parse	that dominates the phrase
# Depth of syntactic phrase	Phrase is a pronoun	Grammatical function, e.g. SUBJ *
# Cardinal numbers	Phrase contains more than 1 DP	Type of pronoun, e.g. demonstrative
# Depth of syntactic phrase	and 1 NP (i.e. phrase contains	Syntactic shape, e.g. apposition with
ignoring unary branching	an embedded argument)	a determiner and attributive modifier
# Apposition phrases	Head noun appears (partly or completely)	Head noun type, e.g. common *
# Year phrases	in previous 10 sentences *	Head noun number, e.g. singular

Table 1: Features of the CRF prediction model (* indicates feature used in baseline model)

CDU parliamentary group may be known to parts of a German audience but not to other people.

We address this by collecting both hearer-known and hearer-unknown definite expressions into one class UNUSED. This does not rule out further subclassification (*known/unknown*) or the possibility of using machine learning techniques to identify this distinction, see Nenkova et al. (2005). The fact that Rahman and Ng (2011) report the highest confusion rate between NEW and MEDIATED entities may have its roots in this issue.

New. Only (specific) indefinites are labeled NEW.

Generic. An issue which is not dealt with in Nissim et al. (2004) are GENERIC expressions as in *Lions have manes*. Reiter and Frank (2010) discuss the task of identifying generic items in a manner similar to the learning tasks presented above, using a Bayesian network. We believe it makes sense to integrate genericity detection into information-status prediction.[3]

4 German data

Our work is based on the DIRNDL radio news corpus of Eckart et al. (2012) which has been hand-annotated with information status labels. We choose a selection of 6668 annotated phrases (1420 sentences). This is an order of magnitude smaller than the annotated Switchboard corpus of Calhoun et al. (2010). We parse each sentence with the German Lexical Functional Grammar of Rohrer and Forst (2006) using the XLE parser in order to automatically extract (morpho-)syntactic and functional features for our model.

5 Prediction Model for Information Status

Cahill and Riester (2009) show that there are asymmetries between pairs of information status labels contained in sentences, i.e. certain classes of expressions tend to precede certain other classes. We therefore treat the prediction of IS labels as a sequence labeling task.[4] We train a CRF using `wapiti` (Lavergne et al., 2010), with the features outlined in Table 1. We also include a basic "coreference" feature, similar to the lexical features of Rahman and Ng (2011), that fires if there is some lexical overlap of nouns (or compound nouns) in the preceding 10 sentences. The original label set described in Riester et al. (2010) contains 21 labels. Here we work with a subset of maximally 12 labels, but also consider smaller subsets of labels and carry out a mapping to the Nissim (2006) label set (Table 2).[5] We run a 10-fold cross-validation experiment and report average prediction accuracy. The results are given in Table 3a. As an informed baseline, we run the same cross-validation experiment with a subset of features that roughly correspond to the features of Nissim (2006). Our models perform statistically significantly better than the baseline ($p < 0.001$, using the approximate randomization test) for all label sets.

[3]Note that in coreference annotation it is an open question whether two identical generic terms should count as coreferent.

[4]Preliminary experimental evidence showed that the CRF performed slightly better than a simple multiclass logistic regression model (e.g. compare 72.19 to 72.43 in Table 3a).

[5]Unfortunately, due to underlying theoretical differences, it is impossible to map between the Riester label set and the extended label set used in Rahman and Ng (2012).

Total	Riester 1	Riester 2	Riester 3	Nissim '06
462	GIVEN-PRONOUN	GIVEN-PRONOUN	GIVEN	OLD
143	GIVEN-REFLEXIVE	GIVEN-REFLEXIVE		
427	GIVEN-EPITHET	GIVEN-NOUN		
169	GIVEN-REPEATED			
204	GIVEN-SHORT			
265	SITUATIVE	SITUATIVE	SITUATIVE	
449	BRIDGING	BRIDGING	BRIDGING	MEDIATED
1271	UNUSED-KNOWN	UNUSED-KNOWN	UNUSED	
1227	UNUSED-UNKNOWN	UNUSED-UNKNOWN		NEW
1282	NEW	NEW	NEW	
632	GENERIC	GENERIC	GENERIC	
96	EXPLETIVE	EXPLETIVE	EXPLETIVE	OTHER

Table 2: Varying the granularity of the label sets

As expected, the less fine-grained a label set, the easier it is to predict the labels. It remains for future work to show the effect of different label set granularities in practical applications. We approximate gold standard coreference information from the manually annotated labels (e.g. all GIVEN label types are by their nature coreferent), and carry out an experiment with gold-standard approximation of coreference marking. These results are also reported in Table 3a. Here we see a clear performance difference in the effect of gold-standard coreference on the Riester label set (increasing around 6-10%), compared to the Nissim label set (decreasing slightly). This is an artifact of the way the mapping was carried out, deriving the gold standard coreference information from the Riester label set. There is not a one-to-one mapping between OLD and GIVEN, and, in the Riester label set, coreferential entities that are labeled as SITUATIVE (deictic terms) are not recognized as such.

The feature set in Table 1 reflects the morpho-syntactic properties of the phrases to be labeled. Sometimes world knowledge is required in order to be able to accurately predict a label; for example, to know that *the pope* can be categorized as UNUSED-KNOWN, because it can occur discourse-initially, whereas *the priest* must usually be categorized as GIVEN. The BRIDGING relationship is also difficult to capture without some world knowledge. For example, to infer that *the waitress* can be categorized as BRIDGING in the context of *the restaurant* requires information that links the two concepts. Rahman and Ng (2012) also note this and include features based on FrameNet, WordNet and the ReVerb corpus for English.

For German, we address this issue by introducing two further types of features into our model based on the GermaNet resource (Hamp and Feldweg, 1997). The first type is based on the GermaNet synset of the head noun in the phrase and its distance from the root node (the assumption is that entities closer to root are more generic than those further away). The second include the sum and maximum of the Lin semantic relatedness measures (Lin, 1998) of how similar the head noun of the phrase is to the other nouns in current and immediately preceding sentence surrounding the phrase (calculated with GermaNet Pathfinder; Finthammer and Cramer, 2008). The results are given in Table 3b. Here we see a consistent increase in performance of around 4% for each label set over the model that does not include the GermaNet features. Again, we see the same decrease in performance on the Nissim label set when using gold standard coreference information.

Label Set	Accuracy	Gold coref.	Baseline feats.
Riester 1	65.49	72.49	57.25
Riester 2	67.21	76.88	58.82
Riester 3	72.43	82.22	64.20
Nissim '06	76.24	74.06	71.70

(a) Only morpho-syntactic features

Label Set	Accuracy	Gold coreference
Riester 1	69.56	76.62
Riester 2	71.99	79.86
Riester 3	75.82	84.76
Nissim '06	79.61	78.46

(b) Morpho-syntactic + GermaNet features

Table 3: Cross validation accuracy results

6 Conclusion

In this paper we presented a model for automatically labeling German text with fine-grained information status labels. The results reported here show that we can achieve high accuracy prediction on a complex text type (transcribed radio news), even with a limited amount of data.

References

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Stefan Baumann and Arndt Riester. to appear. Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta and P. Prieto, editors, *Prosody and Meaning*. Mouton de Gruyter, Berlin.

David L. Bean and Ellen Riloff. 1999. Corpus-Based Identification of Non-Anaphoric Noun Phrases. In *Proceedings of ACL*, pages 373–380, College Park, MD.

Aoife Cahill and Arndt Riester. 2009. Incorporating Information Status into Generation Ranking. In *Proceedings of ACL-IJCNLP*, pages 817–825, Singapore.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-Format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Pascal Denis and Jason Baldridge. 2007. Global Joint Determination of Anaphoricity and Coreference Resolution Usinger Integer Programming. In *Proceedings of ACL-HLT*, Rochester, NY.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In C. Chiarcos et al., editors, *Linked Data in Linguistics*, pages 65–76, Berlin. Springer.

Marc Finthammer and Irene Cramer. 2008. Exploring and Navigating: Tools for GermaNet. In *Proceedings of LREC*, Marrakech, Morocco.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Manfred Krifka. 2007. Basic Notions of Information Structure. In C. Féry and M. Krifka, editors, *The Notions of Information Structure*, pages 57–68. Universitätsverlag Potsdam.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings of ACL*, pages 504–513.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *International Conference on Machine Learning*, pages 296–304.

Ani Nenkova, Advaith Siddharthan, and Kathleen McKeown. 2005. Automatically Learning Cognitive Status for Multi-Document Summarization of Newswire. In *Proceedings of HLT/EMNLP*, pages 241–248, Vancouver.

Vincent Ng and Claire Cardie. 2002. Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. In *Proceedings of COLING*, pages 730–736, Taipei, Taiwan.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An Annotation Scheme for Information Status in Dialogue. In *Proceedings of LREC*, Lisbon.

Malvina Nissim. 2006. Learning Information Status of Discourse Entities. In *Proceedings of EMNLP*, pages 94–102, Sydney.

Massimo Poesio and Renata Vieira. 1998. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, 24(2).

Ellen F. Prince. 1981. Toward a Taxonomy of Given-New Information. In P. Cole, editor, *Radical Pragmatics*, pages 233–255. Academic Press, New York.

Ellen F. Prince. 1992. The ZPG Letter: Subjects, Definiteness and Information Status. In W. Mann and S. Thompson, editors, *Discourse Description*, pages 295–325. Benjamins, Amsterdam.

Altaf Rahman and Vincent Ng. 2011. Learning the Information Status of Noun Phrases in Spoken Dialogues. In *Proceedings of EMNLP*, pages 1069–1080, Edinburgh.

Altaf Rahman and Vincent Ng. 2012. Learning the Fine-Grained Information Status of Discourse Entities. In *Proceedings of EACL 2012*, Avignon, France.

Nils Reiter and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of ACL*, pages 40–49, Uppsala, Sweden.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of LREC*, Valletta, Malta.

Christian Rohrer and Martin Forst. 2006. Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of LREC*, Genoa, Italy.

Roger Schwarzschild. 1999. GIVENness, AvoidF, and other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.

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

Olga Uryupina. 2003. High-precision Identification of Discourse New and Unique Noun Phrases. In *Proceedings of the ACL Student Workshop*, pages 80–86, Sapporo.

Guodong Zhou and Fang Kong. 2011. Learning Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation. *Journal of Computer Science and Technology*, 26(1).