Learning Information Status of Discourse Entities

Malvina Nissim*
Laboratory for Applied Ontology
Institute for Cognitive Science and Technology
National Research Council (ISTC-CNR), Roma, Italy
malvina.nissim@loa-cnr.it

Abstract

In this paper we address the issue of automatically assigning information status to discourse entities. Using an annotated corpus of conversational English and exploiting morpho-syntactic and lexical features, we train a decision tree to classify entities introduced by noun phrases as *old*, *mediated*, or *new*. We compare its performance with hand-crafted rules that are mainly based on morpho-syntactic features and closely relate to the guidelines that had been used for the manual annotation. The decision tree model achieves an overall accuracy of 79.5%, significantly outperforming the hand-crafted algorithm (64.4%). We also experiment with binary classifications by collapsing in turn two of the three target classes into one and retraining the model. The highest accuracy achieved on binary classification is 93.1%.

1 Introduction

Information structure is the way a speaker or writer organises known and new information in text or dialogue. Information structure has been the subject of numerous and very diverse linguistic studies (Halliday, 1976; Prince, 1981; Hajičová, 1984; Vallduví, 1992; Lambrecht, 1994; Steedman, 2000, for instance), thus also yielding a wide range of terms and definitions (see (Vallduví, 1992; Kruijff-Korbayová and Steedman, 2003) for a discussion). In the present study, we adopt the term "Information Status", following the definition employed for the annotation of the corpus we use for our experiments (Nissim et al., 2004). Information status describes to which degree a discourse entity is *available* to the hearer, in terms of the speaker's assumptions about the hearer's knowledge and beliefs. Although there is a fine line in the distinction between Information Status and Information Structure, it is fair to say that whereas the latter models wider discourse coherence, the former focuses mainly on the local level of discourse entities. Section 2 provides more details on how this notion is encoded in our corpus.

Information status has generated large interest among researchers because of its complex interaction with other linguistic phenomena, thus affecting several Natural Language Processing tasks. Since it correlates with word order and pitch accent (Lambrecht, 1994; Hirschberg and Nakatani, 1996), for instance, incorporating knowledge on information status would be helpful for natural language generation, and in particular text-to-speech systems. Stöber and colleagues, for example, ascribe to the lack of such information the lower performance of text-to-speech compared to concept-to-speech generation, where such knowledge could be made directly available to the system (Stöber et al., 2000).

Another area where information status can play an important role is anaphora resolution. A major obstacle in the resolution of definite noun phrases with full lexical heads is that only a small proportion of them is actually anaphoric (ca. 30% (Vieira and Poesio, 2000)). Therefore, in the absence of anaphoricity information, a resolution system will try to find an antecedent also for non-

*The work reported in this paper was carried out while the author was a research fellow at the Institute for Communicating and Collaborative Systems of the University of Edinburgh, United Kingdom, and was supported by a Scottish Enterprise Edinburgh-Stanford Link grant (265000-3102-R36766).

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 94–102,
Sydney, July 2006.

anaphoric definite noun phrases, thus severely affecting performance. There has been recent interest in determining anaphoricity *before* performing anaphora resolution (Ng and Cardie, 2002; Uryupina, 2003), but results have not been entirely satisfactory. Given that old entities are more likely to be referred to by anaphors, for instance, identification of information status could improve anaphoricity determination.

Postolache et al. (2005) have recently shown that learning information structure with high accuracy is feasible for Czech. However, there are yet no studies that explore such a task for English. Exploiting an existing annotated corpus, in this paper we report experiments on learning a model for the automatic identification of information status in English.

2 Data

For our experiments we annotated a portion of the transcribed Switchboard corpus (Godfrey et al., 1992), consisting of 147 dialogues (Nissim et al., 2004).[1] In the following section we provide a brief description of the annotation categories.

2.1 Annotation

Our annotation of information status mainly builds on (Prince, 1992), and employs a distinction into *old*, *mediated*, and *new* entities similar to the work of (Strube, 1998; Eckert and Strube, 2001).

All noun phrases (NPs) were extracted as markable entities using pre-existing parse information (Carletta et al., 2004). An entity was annotated as *new* if it has not been previously referred to and is yet unknown to the hearer. The tag *mediated* was instead used whenever an entity that is newly mentioned in the dialogue can be inferred by the hearer thanks to prior or general context.[2] Typical examples of mediated entities are generally known objects (such as "the sun", or "the Pope" (Löbner, 1985)), and bridging anaphors (Clark, 1975; Vieira and Poesio, 2000), where an entity is related to a previously introduced one. Whenever an entity was neither new nor mediated, it was considered as *old*.

[1] Switchboard is a collection of spontaneous phone conversations, averaging six minutes in length, between speakers of American English on predetermined topics. A third of the corpus is syntactically parsed as part of the Penn Treebank (Marcus et al., 1993)

[2] This type corresponds to Prince's (1981; 1992) *inferrables*.

In order to account for the complexity of the notion of information status, the annotation also includes a sub-type classification for old and mediated entities that provides a finer-grained distinction with information on why a given entity is mediated (e.g., set-relation, bridging) or old (e.g., coreference, generic pronouns). In order to test the feasibility of automatically assigning information status to discourse entities, we took a modular approach and only considered the coarser-grained distinctions for this first study. Information about the finer-grained subtypes will be used in future work.

In addition to the main categories, we used two more annotation classes: a tag *non-applicable*, used for entities that were wrongly extracted in the automatic selection of markables (e.g. "course" in "of course"), for idiomatic occurrences, and expletive uses of "it"; and a tag *not-understood* to be applied whenever an annotator did not fully understand the text. Instances annotated with these two tags, as well as all traces, which were left unannotated, were excluded from all our experiments.

Inter-annotator agreement was measured using the kappa (K) statistics (Cohen, 1960; Carletta, 1996) on 1,502 instances (three Switchboard dialogues) marked by two annotators who followed specific written guidelines. Given that the task involves a fair amount of subjective judgement, agreement was remarkably high. Over the three dialogues, the annotation yielded $K = .845$ for the old/med/new classification ($K = .788$ when including the finer-grained subtype distinction). Specifically, "old" proved to be the easiest to distinguish, with $K = .902$; for "med" and "new" agreement was measured at $K = .800$ and $K = .794$, respectively. A value of $K > .76$ is usually considered good agreement. Further details on the annotation process and corpus description are provided in (Nissim et al., 2004)

2.2 Setup

We split the 147 dialogues into a training, a development and an evaluation set. The training set contains 40,865 NPs distributed over 94 dialogues, the development set consists of 23 dialogues for a total of 10,565 NPs, and the evaluation set comprises 30 dialogues with 12,624 NPs. Instances were randomised, so that occurrences of NPs from the same dialogue were possibly split across the different sets.

Table 1 reports the distribution of classes for the training, development and evaluation sets. The distributions are similar, with a majority of old entities, followed by mediated entities, and lastly by new ones.

Table 1: Information status distribution of NPs in training, development and evaluation sets

	TRAIN	DEV	EVAL
old	19730 (48.3%)	5181 (49.0%)	6049 (47.9%)
med	15184 (37.1%)	3762 (35.6%)	4644 (36.8%)
new	5951 (14.6%)	1622 (15.4%)	1931 (15.3%)
total	40865 (100%)	10565 (100%)	12624 (100%)

3 Classification with hand-crafted rules

The target classes for our classification experiments are the annotation tags: old, mediated, and new. As baseline, we could take a simple "most-frequent-class" assignment that would classify all entities as old, thus yielding an accuracy of 47.9% on the evaluation set (see Table 1). Although the "all-old" assumption makes a reasonable baseline, it would not provide a particularly interesting solution from a practical perspective, since a dialogue should also contain not-old information. Thus, rather than adopting this simple strategy, we developed a more sophisticated baseline working on a set of hand-crafted rules.

This hand-crafted algorithm is based on rather straightforward, intuitive rules, partially reflecting the instructions specified in the annotation guidelines. As shown in Figure 1, the top split is the NP type: whether the instance to classify is a pronoun, a proper noun, or a common noun. The other information that the algorithm uses is about complete or partial string overlapping with respect to the dialogue's context. For common nouns we also consider the kind of determiner (*def*inite, *indef*inite, *dem*onstrative, *poss*essive, or *bare*).

In order to obtain the NP type information, we exploited the pre-existing morpho-syntactic treebank annotation of Switchboard. Whenever the extraction failed, we assigned a type "other" and always backed-off these cases to old (the most frequent class in training data). Values for the other features were obtained by simple pattern matching and NP extraction.

Evaluation measures The algorithm's performance is evaluated with respect to its general accuracy (Acc): the number of correctly classified instances over all assignments. Moreover, for each class (c), we report precision (P), recall (R), and f-score (F) thus calculated:

$$\mathrm{P}_c = \frac{correct\ assignments\ of\ c}{total\ assignments\ of\ c}$$

$$\mathrm{R}_c = \frac{correct\ assignments\ of\ c}{total\ corpus\ instances\ of\ c}$$

$$\mathrm{F}_c = \frac{2P_cR_c}{P_c+R_c}$$

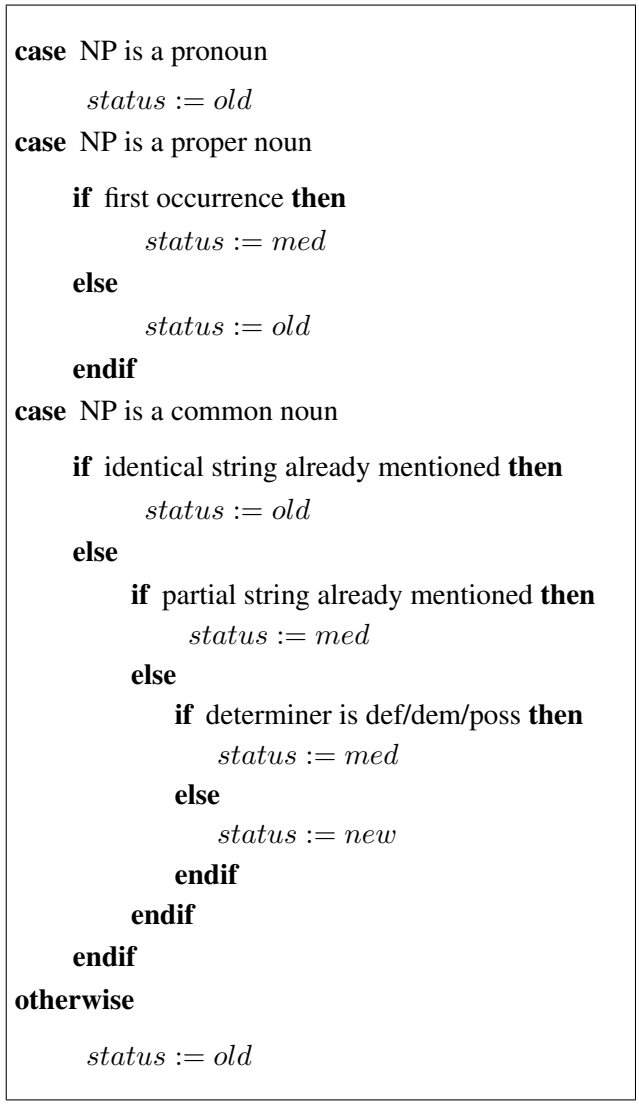

Figure 1: Hand-crafted rule-based algorithm for the assignment of information status to NPs.

The overall accuracy of the rule-based algorithm is 65.8%. Table 2 shows the results for each target class in both the development and evaluation sets. We discuss results on the latter.

Although a very high proportion of old entities is correctly retrieved (93.5%), this is done with relatively low precision (66.7%). Moreover, both precision and recall for the other classes are disappointing. Unsurprisingly, the rules that apply to common nouns (the most ambiguous with respect to information status) generate a large num-

Table 2: Per class performance of hand-crafted rules on the development and evaluation sets

	DEV			EVAL		
	P	R	F	P	R	F
old	.677	.932	.784	.667	.935	.779
med	.641	.488	.554	.666	.461	.545
new	.517	.180	.267	.436	.175	.250

ber of false positives. The rule that predicts an old entity in case of a full previous mention, for example, has a precision of only 39.8%. Better, but not yet satisfactory, is the precision of the rule that predicts a mediated entity for a common noun that has a previous partial mention (64.7%). The worst performing rule is the one that assigns the most frequent class (old) to entities of syntactic type "other", with a precision of 35.4%. To give an idea of the correlation between NP type and information status, in Table 3 we report the distribution observed in the evaluation set.

Table 3: Distribution of information status over NP types in the evaluation set

	old	med	new
pronoun	4465	159	13
proper	107	198	27
common	752	2874	1256
other	725	1413	635

4 Learning Information Status

Our starting point for the automatic assignment of information status are the three already introduced classes: old, mediated and new. Additionally, we experiment with binary classifications, by collapsing mediated entities in turn with old and new ones.

For training, developing and evaluating the model we use the split described in Section 2.2 (see Table 1). Performance is evaluated according to overall accuracy and per class precision, recall, and f-score as described in Section 3. To train a C4.5 decision tree model we use the J48 Weka implementation (Witten and Frank, 2000). The choice of features to build the tree is described in the following section.

4.1 Features

The seven features we use are automatically extracted from the annotated data exploiting pre-existing morpho-syntactic markup and using simple pattern matching techniques. They are summarised in Table 4.

Table 4: Feature set for learning experiments

FEATURE	VALUES
full prev mention	numeric
mention time	{first,second,more}
partial prev mention	{yes,no,na}
determiner	{bare,def,dem,indef,poss,na}
NP length	numeric
grammatical role	{subject,subjpass,object,pp,other}
NP type	{pronoun,common,proper,other}

The choice of features is motivated by the following observations. The information coming from partial previous mentions is particularly useful for the identification of mediated entities. This should account specifically for cases of mediation via set-relations; for example, "your children" would be considered a partial previous mention of "my children" or "your four children". The value "na" stands for "non-applicable" and is mainly used for pronouns. Full previous mention is likely to be a good indicator of old entities. Both full and partial previous mentions are calculated within each dialogue without any constraints based on distance.

NP type and determiner type are expected to be helpful for all categories, with pronouns, for instance, tending to be old and indefinite NPs being often new. We included the length of NPs (measured in number of words) since linguistic studies have shown that old entities tend to be expressed with less lexical material (Wasow, 2002). In experiments on the development data we also included the NP string itself, on the grounds that it might be of use in cases of general mediated instances (common knowledge entities), such as "the sun", "people", "Mickey Mouse", and so on. However, this feature turned out to negatively affect performance, and was not included in the final model.

4.2 Results

With an overall final accuracy of 79.5% on the evaluation set, C4.5 significantly outperforms the hand-crafted algorithm (65.8%). Although the identification of old entities is quite successful (F=.928), performance is not entirely satisfactory. This is especially true for the classification of new entities, for which the final f-score is .320, mainly due to extremely low recall (.223). Mediated entities, instead, are retrieved with a fairly low precision but higher recall. Table 5 summarises precision, recall, and f-score for each class.

Table 5: Per class performance of C4.5 on the development and evaluation sets

	DEV			EVAL		
	P	R	F	P	R	F
old	.935	.911	.923	.941	.915	.928
med	.673	.878	.762	.681	.876	.766
new	.623	.234	.341	.563	.223	.320

The major confusion in the classification arises between mediated and new (the most difficult decision to make for human annotators too, see Section 2.1), which are often distinguished on the basis of world knowledge, not available to the classifier. This is clearly shown by the confusion matrix in Table 6: the highest proportion of mistakes is due to 1,453 new instances classified as mediated. Also significant is the wrong assignment of mediated tags to old entities. Such behaviour of the classifier is to be expected, given the 'in-between' nature of mediated entities.

Table 6: Confusion matrix for evaluation set. C=Classifier tag; G=Gold tag

C → G ↓	old	med	new
old	5537	452	60
med	303	4066	275
new	47	1453	431

4.3 Classification with two categories only

Given the above observations, we collapsed mediated entities in turn with old ones (focusing on their non-newness) or new ones (enhancing their non complete givenness), thus reducing the task to a binary classification.

Since it appears to be more difficult to distinguish mediated and new rather than mediated and old (Table 6), we expect the classifier to perform better when mediated is binned with new rather than old. Also, in the case where mediated and old entities are collapsed into one single class as opposed to new ones, the distribution of classes becomes highly skewed towards old entities (84.7%) so that the learner is likely to lack sufficient information for identifying new entities.

Table 7 shows the final accuracy for the two binary classifications (and the three-way one). As expected, when mediated entities are joint with new ones, the classifier performs best (93.1%), with high f-scores for both old and new, and is significantly better than the alternative binary classification (t-test, $p < 0.001$). Indeed, the old+med vs new classification is nearly an all-old assignment and its overall final accuracy (85.5%) is not a significant improvement over the all-old baseline (84.7%). Results suggest that mediated NPs are more similar to new than to old entities and might provide interesting feedback for the theoretical assumptions underlying the corpus annotation.

4.4 Comparison with two categories only

For a fair comparison, we performed a two-way classification using the hand-crafted algorithm, which had to be simplified to account for the lack of a mediated class.

In the case where all mediated instances where collapsed together with the old ones, the decision rules are very simple: pronouns, proper nouns, and common nouns that have been previously fully or partially mentioned are classified as old; first mention common nouns are new; everything else is old. Both precision and recall for old instances are quite high (.868 and .906 respectively), for a resulting f-score of .887. Conversely, the performance on identifying new entities is very poor, with a precision of .337 and a recall of .227, for a combined f-score of .271. The overall accuracy is .803, and this is significantly lower than the performance of C4.5, which achieves an overall accuracy of .850 (t-test, $p < 0.001$).

When mediated entities are collapsed with new ones, rule-based classification is done again with a very basic algorithm derived from the rules in Figure 1: pronouns are old; proper nouns are new if first mention, old otherwise; common nouns that have been fully previously mentioned are old, otherwise new. Everything else is new, which in the training set is now the most frequent class (51.7%). The overall accuracy of .849 is significantly lower than that achieved by C4.5, which is .931 (t-test, $p < 0.001$). Differently from the previous case (mediated collapsed with old), the performance on each class is comparable, with a precision, recall and f-score of .863, .815, and .838 for old and of .838, .881, and .859 for new.

5 Discussion

5.1 Influence of training size

In order to assess the contribution of training size to performance, we experimented with increas-

Table 7: Overview of accuracy for hand-crafted rules and C4.5 on three-way and binary classifications on development and evaluation sets

	DEV		EVAL	
classification	rules	C4.5	rules	C4.5
old vs med vs new	.658	.796	.644	.795
old+med vs new	.810	.861	.803	.855
old vs med+new	.844	.926	.849	.931

ingly larger portions of the training data (from 50 to 30,000 instances). For each training size we ran the classifier 5 times, each with a different randomly picked set of instances. This was done for the three-way and the two binary classifications. Reported results are always averaged over the 5 runs. Figure 2 shows the three learning curves.

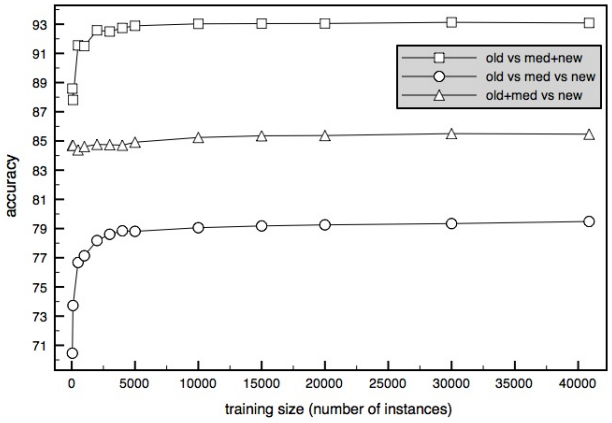

Figure 2: Learning curves for three- and two-way classifications

The curve for the three-way classification shows a slight constant improvement, though it appears to reach a plateau after 5,000 instances. The result obtained training on the full set (40865 instances) is significantly better only if compared to a training set of 4,000 or less (t-test, $p < 0.05$). No other significant difference in accuracy can be observed.

Increasing the training size over 5,000 instances when learning to classify old+mediated vs new leads to a slight improvement due to the learner being able to identify some new entities. With a smaller training set the proportion of new entities is far too small to be of use. However, as said, the overall final accuracy of 85.5% (see Table 7) does not significantly improve over the baseline.

Table 8: Performance of leave-one-out and single-feature classifiers on three-way classification

FEATURE	ACCURACY	
	removed	single
full prev mention	.793	.730
mention time	.795	.730
partial prev mention	.791	.769
determiner	.789	.775
NP length	.793	.733
gram role	.782	.656
NP type	.784	.701
full set	.795	

5.2 Feature contribution

We are also interested in the contribution of each single feature. Therefore, we ran the classifier again, leaving out one feature at a time. No significant drop or gain was observed in any of the runs (t-test, $p < 0.01$), though the worst detriments were yielded by removing the grammatical role and the NP type. These two features, however, also appear to be the least informative in single-feature classification experiments, thus suggesting that such information comes very useful only when combined with other evidence (see also Section 5.4. All results for leave-one-out and single-feature classifiers are shown in Table 8.

5.3 Error Analysis

The overwhelming majority of mistakes (1,453, 56.1% of all errors) in the three-way classification stems from classifying as mediated entities that are in fact new (Table 6). Significant confusion arises from proper nouns, as they are annotated as mediated or new entities, depending on whether they are generally known (such as names of US presidents, for example), or domain/community-specific (such as the name of a local store that only the speaker knows). This inconsistency in the annotation might reflect well the actual status of entities in the dialogues, but it can be misleading for the classifier.

Another large group of errors is formed by old entities classified as mediated (452 cases). This is probably due to the fact that the first node in the decision tree is the "partial mention" feature (see Figure 3). The tree correctly captures the fact that a firstly mentioned entity which has been partially mentioned before is mediated. An entity that has a previous partial mention but also a full previous mention is classified as old only if it is a proper noun or a pronoun, but as mediated if it is a common noun. This yields a large number of mis-

takes, since many common nouns that have been previously mentioned (both in full and partially) are in fact old. Another problem with previous mentions is the lack of restriction in distance: we consider a previous mention any identical mention of a given NP anywhere in the dialogue, and we have no means of checking that it is indeed the same entity that is referred to. A way to alleviate this problem might be exploiting speaker turn information. Using anaphoric chains could also be of help, but see Section 6.

5.4 Learnt trees meet hand-crafted rules

The learnt trees provide interesting insights on the intuitions behind the choice of hand-crafted rules.

```
partial = yes
|   full <= 1
|   |   det = def: med
|   |   det = indef
|   |   |   length <= 2
|   |   |   |   gramm = subj: med
|   |   |   |   gramm = subjpassive: new
|   |   |   |   gramm = obj: med
|   |   |   |   gramm = pp: med
|   |   |   |   gramm = other
|   |   |   |   |   type = proper: med
|   |   |   |   |   type = common: new
|   |   |   |   |   type = pronoun: new
|   |   |   |   |   type = other: med
|   |   |   length > 2: med
|   |   det = dem
|   |   |   gramm = subj
. . .
```

Figure 3: Top of C.5, full training set, three classes

Figure 3 shows the top of C4.5 (trained on the full training set for the three-way classification), which looks remarkably different from the rules in Figure 1. We had based our decision of emphasising the importance of the NP type on the linguistic evidence that different syntactic realisations reflect different degrees of availability of discourse entities (Givón, 1983; Ariel, 1990; Grosz et al., 1995). In the learnt model, however, knowledge about NP type is only used as subordinate to other features. This is indeed mirrored in the fact that removing NP type information from the feature set causes accuracy to drop, but a classifier building on NP type alone performs poorly (see Table 8).[3] Interestingly, though, more informative knowledge about syntactic form seems to be derived from the determiner type, which helps distinguish degrees of oldness among common nouns.

[3]The NPtype-only classifier assigns old to pronouns and med to all other types; it never assigns new.

5.5 Naive Bayes model

For additional comparison, we also trained a Naive Bayes classifier with the same experimental settings. Results are significantly worse than C4.5's in all three scenarios (t-test, $p < 0.005$), with an accuracy of 74.6% in the three-way classification, 63.3% for old+mediated vs new, and 91.0% for old vs mediated+new. The latter distribution appears again to be the easiest to learn.

6 Related Work

To our knowledge, there are no other studies on the automatic assignment of information status in English. Recently, (Postolache et al., 2005) have reported experiments on learning information structure in the Prague TreeBank. The Czech treebank is annotated following the Topic-Focus articulation theory (Hajičová et al., 1998). The theoretical definitions underlying the Prague Treebank and the corpus we are using are different, with the former giving a more global picture of information structure, and the latter a more entity-specific one. For this reason, and due to the fact that Postolache et al.'s experiments are on Czech (with a freer word order than English), comparing results is not straightforward.

Their best system (C4.5 decision tree) achieves an accuracy of 90.69% on the topic/focus identification task. This result is comparable with the result we obtain when training and testing on the corpus where mediated and new entities are not distinguished (93.1%). Postolache and colleagues also observe a slowly flattening learning curve after a very small amount of data (even 1%, in their case). Therefore, they predict an increase in performance will mainly come from better features rather than more training data. This is likely to be true in our case as well, also because our feature set is currently small and we will further benefit from incorporating additional features. Postolache et al. use a larger feature set, which also includes coreference information. The corpus we use has manually annotated coreference links. However, because we see anaphoricity determination as a task that could benefit from automatic information status assignment, we decided not to exploit this information in the current experiments. Moreover, we did not want our model to rely too heavily on a feature that is not easy to obtain automatically.

7 Conclusions and Future Work

We have presented a model for the automatic assignment of information status in English. On the three-way classification into old, mediated, and new that reflects the corpus annotation tags, the learnt tree outperforms a hand-crafted algorithm and achieves an accuracy of 79.5%, with high precision and recall for old entities, high recall for mediated entities, and a fair precision, but very poor recall, for new ones. When we collapsed mediated and new entities into one category only opposing this to old ones, the classifier performed with an accuracy of 93.1%, with high f-scores for both classes. Binning mediated and old entities together did not produce interesting results, mainly due to the highly skewed distribution of the resulting corpus towards old entities. This suggests that mediated entities are more similar to new than to old ones, and might provide interesting feedback for the theoretical assumptions underlying the annotation. Future work will examine specific cases and investigate how such insights can be used to make the theoretical framework more accurate.

As the first experiments run on English to learn information status, we wanted to concentrate on the task itself and avoid noise introduced by automatic processing. More realistic settings for integrating an information status model in a large-scale NLP system would imply obtaining syntactic information via parsing rather than directly from the treebank. Future experiments will assess the impact of automatic preprocessing of the data.

Results are very promising but there is room for improvement. First, the syntactic category "other" is far too large, and finer distinctions must be made by means of better extraction rules from the trees. Second, and most importantly, we believe that using more features will be the main trigger of higher accuracy. In particular, we plan to use additional lexical and relational features derived from knowledge sources such as WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 1998) which should be especially helpful in distinguishing mediated from new entities, the most difficult decision to make. For example, an entity that is linked in WordNet (within a given depth) and/or FrameNet to a previously introduced one is more likely to be mediated than new.

Additionally, we will attempt to exploit dialogue turns, since knowing which speaker said what is clearly very valuable information. In a similar vein, we will experiment with distance measures, in terms of turns, sentences, or even time, for determining when an introduced entity might stop to be available.

We also plan to run experiments on the automatic classification of old and mediated subtypes (the finer-grained classification) that is included in the corpus but that we did not consider for the present study (see Section 2.1). The major benefit of this would be a contribution to the resolution of bridging anaphora.

References

Mira Ariel. 1990. *Accessing Noun Phrase Antecedents*. Routledge, London-New York.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of COLING-ACL*, pages 86–90.

Jean Carletta, Shipra Dingare, Malvina Nissim, and Tatiana Nikitina. 2004. Using the NITE XML Toolkit on the Switchboard Corpus to study syntactic choice: a case study. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004), Lisbon, May 2004.*

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Herbert H. Clark. 1975. Bridging. In Roger Schank and Bonnie Nash-Webber, editors, *Theoretical Issues in Natural Language Processing*. The MIT Press, Cambridge, MA.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46.

Miriam Eckert and Michael Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Talmy Givón. 1983. Introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A Quantitative Cross-language Study*. John Benjamins, Amsterdam/Philadelphia.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, pages 517–520.

Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Eva Hajičová, Barbara Partee, and Petr Sgall. 1998. Topic-focus articulation, tripartite structures, and semantic content. In *Studies in Linguistics and Philosophy*, volume 71. Dordrecht.

Eva Hajičová. 1984. Topic and focus. In Petr Sgall, editor, *Contributions to Functional Syntax. Semantics and Language Comprehension (LLSEE 16)*, pages 189–202. John Benjamins, Amsterdam.

M.A.K. Halliday. 1976. Notes on transitivity and theme in English. Part 2. *Journal of Linguistics*, 3(2):199–244.

Julia Hirschberg and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction giving monologues. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*.

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

Knud Lambrecht. 1994. *Information structure and sentence form. Topic, focus, and the mental representation of discourse referents*. Cambridge University Press, Cambridge.

Sebastian Löbner. 1985. Definites. *Journal of Semantics*, 4:279–326.

Mitchell Marcus, Beatrice Santorini, and May Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. *Computational Linguistics*, 19:313–330.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proc of the 19th International Conference on Computational Linguistics; Taipei, Taiwan*, pages 730–736.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004), Lisbon, May 2004.*

Oana Postolache, Ivana Kruijff-Korbayova, and Geert-Jan Kruijff. 2005. Data-driven approaches for information structure identification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*. Academic Press, New York.

Ellen Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In Sandra Thompson and William Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, Philadelphia/Amsterdam.

Mark Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge, MA.

K. Stöber, P. Wagner, Jörg Helbig, S. Köster, D. Stall, M. Thomas, J. Blauert, W. Hess, R. Hoffmann, and H. Mangold. 2000. Speech synthesis using multilevel selection and concatenation of units from large speech corpora. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 519–534. Springer-Verlag, Berlin.

Michael Strube. 1998. Never look back: An alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 1251–1257, Montréal, Québec, Canada.

Olga Uryupina. 2003. High-precision identification of discourse new and unique noun phrases. In *Proc. of the ACL 2003 Student Workshop*, pages 80–86.

Enric Vallduví. 1992. *The Informational Component*. Garland, New York.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4).

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Diego, CA.