Collective Classification for Fine-grained Information Status

Katja Markert[1,2], **Yufang Hou**[2], **Michael Strube**[2]
[1] School of Computing, University of Leeds, UK, scskm@leeds.ac.uk
[2] Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany
(yufang.hou|michael.strube)@h-its.org

Abstract

Previous work on classifying information status (Nissim, 2006; Rahman and Ng, 2011) is restricted to coarse-grained classification and focuses on conversational dialogue. We here introduce the task of classifying fine-grained information status and work on written text. We add a fine-grained information status layer to the Wall Street Journal portion of the OntoNotes corpus. We claim that the information status of a mention depends not only on the mention itself but also on other mentions in the vicinity and solve the task by collectively classifying the information status of all mentions. Our approach strongly outperforms reimplementations of previous work.

1 Introduction

Speakers present already known and yet to be established information according to principles referred to as *information structure* (Prince, 1981; Lambrecht, 1994; Kruijff-Korbayová and Steedman, 2003, inter alia). While information structure affects all kinds of constituents in a sentence, we here adopt the more restricted notion of *information status* which concerns only discourse entities realized as noun phrases, i.e. *mentions*[1]. Information status (*IS* henceforth) describes the degree to which a discourse entity is available to the hearer with regard to the speaker's assumptions about the hearer's knowledge and beliefs (Nissim et al., 2004). `Old` mentions are known to the hearer and have been referred to previously. `Mediated` mentions have not been mentioned before but are also not autonomous, i.e., they can only be correctly interpreted by reference to another mention or to prior world knowledge. All other mentions are `new`.

IS can be beneficial for a number of NLP tasks, though the results have been mixed. Nenkova et al. (2007) used IS as a feature for generating pitch accent in conversational speech. As IS is restricted to noun phrases, while pitch accent can be assigned to any word in an utterance, the experiments were not conclusive. For determining constituent order of German sentences, Cahill and Riester (2009) incorporate features modeling IS to good effect. Rahman and Ng (2011) showed that IS is a useful feature for coreference resolution.

Previous work on learning IS (Nissim, 2006; Rahman and Ng, 2011) is restricted in several ways. It deals with conversational dialogue, in particular with the corpus annotated by Nissim et al. (2004). However, many applications that can profit from IS concentrate on written texts, such as summarization. For example, Siddharthan et al. (2011) show that solving the IS subproblem of whether a person proper name is already known to the reader improves automatic summarization of news. Therefore, we here model IS in written text, creating a new dataset which adds an IS layer to the already existing comprehensive annotation in the OntoNotes corpus (Weischedel et al., 2011). We also report the first results on fine-grained IS classification by modelling further distinctions within the category of `mediated` mentions, such as comparative and bridging anaphora (see Examples 1 and 2, re-

[1] Since not all noun phrases are referential, we call noun phrases which carry information status *mentions*.

spectively).[2] Fine-grained IS is a prerequisite to full bridging/comparative anaphora resolution, and therefore necessary to fill gaps in entity grids (Barzilay and Lapata, 2008) based on coreference only. Thus, Examples 1 and 2 do not exhibit any coreferential entity coherence but coherence can be established when the comparative anaphor *others* is resolved to *others than freeway survivor Buck Helm,* and the bridging anaphor *the streets* is resolved to *the streets of Oranjemund*, respectively.

(1) the condition of *freeway survivor Buck Helm* . . . , improved, hospital officials said. Rescue crews, however, gave up hope that **others** would be found.

(2) *Oranjemund, the mine headquarters*, is a lonely corporate oasis of 9,000 residents. Jackals roam **the streets** at night . . .

We approach the challenge of modeling IS via collective classification, using several novel linguistically motivated features. We reimplement Nissim's (2006) and Rahman and Ng's (2011) approaches as baselines and show that our approach outperforms these by a large margin for both coarse- and fine-grained IS classification.

2 Related Work

IS annotation schemes and corpora. We enhance the approach in Nissim et al. (2004) in two major ways (see also Section 3.1). First, comparative anaphora are not specifically handled in Nissim et al. (2004) (and follow-on work such as Ritz et al. (2008) and Riester et al. (2010)), although some of them might be included in their respective `bridging` subcategories. Second, we apply the annotation scheme reliably to a new genre, namely news. This is a non-trivial extension: Ritz et al. (2008) applied a variation of the Nissim et al. (2004) scheme to a small set of 220 NPs in a German news/commentary corpus but found that reliability then dropped significantly to the range of $\kappa = 0.55$ to 0.60. They attributed this to the higher syntactic complexity and semantic vagueness in the commentary corpus. Riester et al. (2010) annotated a German news corpus marginally reliable ($\kappa = 0.66$) for their overall scheme but their confusion matrix shows even lower reliability for several subcategories, most importantly deixis and bridging.

While standard coreference corpora do not contain IS annotation, some corpora annotated for bridging are emerging (Poesio, 2004; Korzen and Buch-Kromann, 2011) but they are (i) not annotated for comparative anaphora or other IS categories, (ii) often not tested for reliability or reach only low reliability, (iii) often very small (Poesio, 2004).

To the best of our knowledge, we therefore present the first English corpus reliably annotated for a wide range of IS categories as well as full anaphoric information for three main anaphora types (coreference, bridging, comparative).

Automatic recognition of IS. Vieira and Poesio (2000) describe heuristics for processing definite descriptions in news text. As their approach is restricted to definites, they only analyse a subset of the mentions we consider carrying IS. Siddharthan et al. (2011) also concentrate on a subproblem of IS only, namely the hearer-old/hearer-new distinctions for person proper names.

Nissim (2006) and Rahman and Ng (2011) both present algorithms for IS detection on Nissim et al.'s (2004) Switchboard corpus. Both papers treat IS classification as a local classification problem whereas we look at dependencies between the IS status of different mentions, leading to collective classification. In addition, they only distinguish the three main categories `old`, `mediated` and `new`. Finally, we work on news corpora which poses different problems from dialogue.

Anaphoricity determination (Ng, 2009; Zhou and Kong, 2009) identifies many or most `old` mentions. However, no distinction between `mediated` and `new` mentions is made. Most approaches to bridging resolution (Meyer and Dale, 2002; Poesio et al., 2004) or comparative anaphora (Modjeska et al., 2003; Markert and Nissim, 2005) address only the selection of the antecedent for the bridging/comparative anaphor, not its recognition. Sasano and Kurohashi (2009) do also tackle bridging recognition, but they depend on language-specific non-transferrable features for Japanese.

[2] All examples in this paper are from the OntoNotes corpus. The mention in question is typed in boldface; antecedents, where applicable, are displayed in italics.

3 Corpus Creation

3.1 Annotation Scheme

Our scheme follows Nissim et al. (2004) in distinguishing three major IS categories `old`, `new` and `mediated`. A mention is `old` if it is either coreferential with an already introduced entity or a generic or deictic pronoun. We follow the OntoNotes (Weischedel et al., 2011) definition of coreference to be able to integrate our annotations with it. This definition includes coreference with noun phrase as well as verb phrase antecedents[3].

`Mediated` refers to entities which have not yet been introduced in the text but are inferrable via other mentions or are known via world knowledge. We distinguish the following six subcategories: The category `mediated/comparative` comprises mentions compared via either a contrast or similarity to another one (see Example 1). This category is novel in our scheme. We also include a category `mediated/bridging` (see Examples 2, 3 and 4). Bridging anaphora can be any noun phrase and are not limited to definite NPs as in Poesio et al. (2004), Gardent and Manuélian (2005), Riester et al. (2010). In contrast to Nissim et al. (2004), antecedents for both comparative and bridging categories are annotated and can be noun phrases, verb phrases or even clauses. The category `mediated/knowledge` is inspired by the hearer-old distinction introduced by Prince (1992) and covers entities generally known to the hearer. It includes many proper names, such as *Poland*.[4] Mentions that are syntactically linked via a possessive relation or a PP modification to other, `old` or `mediated` mentions fall into the type `mediated/synt` (see Examples 5 and 6).[5] With no change to Nissim et al.'s scheme, coordinated mentions where at least one element in the conjunction is `old` or `mediated` are covered by the category `mediated/aggregate`, and mentions referring to a value of a previously mentioned function by the type `mediated/func`.

All other mentions are annotated as `new`, including most generics as well as newly introduced, specific mentions such as Example 7.

(3) Initial steps were taken at *Poland's first environmental conference, which I attended last month*. ... it was no accident that **participants** urged the free flow of information

(4) The Bakersfield supermarket *went out of business* last May. **The reason** was ...

(5) One Washington couple sold *their liquor store*

(6) *the main artery into San Francisco*

(7) the owner was murdered by *robbers*

3.2 Agreement Study

We carried out an agreement study with 3 annotators, of which Annotator A was the scheme developer and first author of this paper. All texts used were from the Wall Street Journal (WSJ) portion of OntoNotes. There were no restrictions on which texts to include apart from (i) exclusion of letters to the editor as they contain cross-document links and (ii) a preference for longer texts with potentially richer discourse structure.

Mentions were automatically preselected for the annotators using the gold-standard syntactic annotation.[6] The existing coreference annotation was automatically carried over to the IS task by marking all mentions in a coreference chain (apart from the first mention in the chain) as `old`. The annotation task consisted of marking all mentions for their IS (`old`, `mediated` or `new`) as well as marking `mediated` subcategories (see Section 3.1) and the antecedents for comparative and bridging anaphora.

The scheme was developed on 9 texts, which were also used for training the annotators. Inter-annotator agreement was measured on 26 new texts, which included 5905 pre-marked potential mentions. The annotations of 1499 of these were carried over from OntoNotes, leaving 4406 potential mentions for annotation and agreement measurement. In addition to

[3] In contrast to Nissim et al. (2004), but in accordance with OntoNotes, we do not consider generics for coreference.

[4] This class corresponds roughly to Nissim et al.'s (2004) mediated/general.

[5] This class expands Nissim et al.'s (2004) `poss` category that only considers possessives but not PP modification.

[6] Some non-mentions such as idioms could not be filtered out via the syntactic annotation and had to be excluded during human annotation.

	A-B	A-C	B-C
Overall Percentage coarse	87.5	86.3	86.5
Overall κ coarse	77.3	75.2	74.7
Overall Percentage fine	86.6	85.3	85.7
Overall κ fine	80.1	77.7	77.3

Table 1: Agreement Results

	A-B	A-C	B-C
κ Non-mention	81.5	78.9	86.0
κ Old	80.5	83.2	79.3
κ New	76.6	74.0	74.3
κ Mediated/Knowledge	82.1	78.4	74.1
κ Mediated/Synt	88.4	87.8	87.6
κ Mediated/Aggregate	87.0	85.4	86.0
κ Mediated/Func	6.0	83.2	6.9
κ Mediated/Comp	81.8	78.3	81.2
κ Mediated/Bridging	70.8	60.6	62.3

Table 2: Agreement Results for individual categories

percentage agreement, we measured Cohen's κ (Artstein and Poesio, 2008) between all 3 possible annotator pairings. We also report single-category agreement for each category, where all categories but one are merged and then κ is computed as usual. Table 1 shows agreement results for the overall scheme at the coarse-grained (4 categories: *non-mention, old, new, mediated*) and the fine-grained level (9 categories: *non-mention, old, new* and the 6 *mediated* subtypes). The results show that the scheme is overall reliable, with not too many differences between the different annotator pairings.[7]

Table 2 shows the individual category agreement for all 9 categories. We achieve high reliability for most categories.[8] Particularly interesting is the fact that hearer-old entities (`mediated/knowledge`) can be identified reliably although all annotators had substantially different backgrounds. The reliability of the category `bridging` is more annotator-dependent, although still higher, sometimes considerably, than other previous attempts at bridging annotation (Poesio et al., 2004; Gardent and Manuélian, 2005; Riester et al., 2010).

3.3 Gold Standard

Our final gold standard corpus consists of 50 texts from the WSJ portion of the OntoNotes corpus- The corpus will be made publically available as OntoNotes annotation layer via `http://www.h-its.org/nlp/download`.

Disagreements in the 35 texts used for annotator training (9 texts) and testing (26 texts) were resolved via discussion between the annotators. An additional 15 texts were annotated by Annotator A. Finally, Annotator A carried out consistency checks over all texts. – The gold standard includes 10,980 true mentions (see Table 3).

Texts		50
Mentions		10,980
old		3237
	coref	3,143
	generic_deictic_pr	94
mediated		3,708
	world knowledge	924
	syntactic	1,592
	aggregate	211
	func	65
	comparative	253
	bridging	663
new		4,035

Table 3: Gold Standard Distribution

4 Features

In this Section, we describe both the local as well as the relational features we use.

4.1 Features for Local Classification

We use the following local features, including the features in Nissim (2006) and Rahman and Ng (2011) to be able to gauge how their systems fare on our corpus and as a comparison point for our novel collective classification approach.

The features developed by Nissim (2006) are shown in Table 4. Nissim shows clearly that these features are useful for IS classification. Thus, subjects are more likely to be `old` as assumed by, e.g., centering theory (Grosz et al.,

[7] Often, annotation is considered highly reliable when κ exceeds 0.80 and marginally reliable when between 0.67 and 0.80 (Carletta, 1996). However, the interpretation of κ is still under discussion (Artstein and Poesio, 2008).

[8] The low reliability of the rare category `func`, when involving Annotator B, was explained by Annotator B forgetting about this category after having used it once. Pair A-C achieved high reliability (κ 83.2 for pair A-C).

Feature	Value
full prev mention	{yes, no, NA}[9]
mention time	{first, second, more}
partial prev mention	{yes, no, NA}
determiner	{bare, def, dem, indef, poss, NA}
NP type	{pronoun, common, proper, other}
NP length	numeric
grammatical role	{subject, subjpass, pp, other}

Table 4: Nissim's (2006) feature set

1995). Also, previously unmentioned proper names are more likely to be hearer-old and therefore `mediated/knowledge`, although their exact status will depend on how well known a particular proper name is.

Rahman and Ng (2011) add all *unigrams* appearing in any mention in the training set as features. They also integrated (via a convolution tree-kernel SVM (Collins and Duffy, 2001)) partial parse trees that capture the generalised syntactic context of a mention e and include the mention's parent and sibling nodes without lexical leaves. However, they use no structure underneath the mention node e itself, assuming that "any NP-internal information has presumably been captured by the flat features".

To these feature sets, we add a small set of other local features *otherlocal*. These track partial previous mentions by also counting partial previous mention time as well as the previous mention of content words only. We also add a mention's number as one of singular, plural or unknown, and whether the mention is modified by an adjective. Another feature encapsulates whether the mention is modified by a comparative marker, using a small set of 10 markers such as *another, such, similar* ... and the presence of adjectives or adverbs in the comparative. Finally, we include the mention's semantic class as one of 12 coarse-grained classes, including location, organisation, person and several classes for numbers (such as date, money or percent).

4.2 Relations for Collective Classification

Both Nissim (2006) and Rahman and Ng (2011) classify each mention individually in a standard supervised ML setting, not considering potential dependencies between the IS categories of different mentions. However, collective or joint classification has made substantial impact in other NLP tasks, such as opinion mining (Pang and Lee, 2004; Somasundaran et al., 2009), text categorization (Yang et al., 2002; Taskar et al., 2002) and the related task of coreference resolution (Denis and Baldridge, 2007). We investigate two types of relations between mentions that might impact on IS classification.

Syntactic parent-child relations. Two `mediated` subcategories account for accessibility via syntactic links to another `old` or `mediated` mention: `mediated/synt` is used when at least one child of a mention is `mediated` or `old`, with child relations restricted to pre- or postnominal possessives as well as PP children in our scheme (see Section 3.1). `mediated/aggregate` is for coordinations in which at least one of the children is `old` or `mediated`. In these two cases, a mention's IS depends directly on the IS of its children. We therefore link a mention m_1 to a mention m_2 via a *hasChild* relation if (i) m_2 is a possessive or prepositional modification of m_1, or (ii) m_1 is a coordination and m_2 is one of its children.

Using such a relational feature catches two birds with one stone: firstly, it integrates the internal structure of a mention into the algorithm, which Rahman and Ng (2011) ignore; secondly, it captures dependencies between parent and child classification, which would not be possible if we integrated the internal structure via flat features or additional tree kernels. We hypothesise that the higher syntactic complexity of our news genre (14.5% of all mentions are `mediated/synt`) will make this feature highly effective in distinguishing between `new` and `mediated` categories.

Syntactic precedence relations. IS is said to influence word order (Birner and Ward, 1998; Cahill and Riester, 2009) and this fact has been exploited in work on generation (Prevost, 1996; Filippova and Strube, 2007; Cahill and Riester, 2009). Therefore, we integrate dependencies between the IS classification of mentions in precedence relations.

m_1 precedes m_2 if (i) m_1 and m_2 are in the same clause, allowing for trace subjects in gerund and infinitive constructions, (ii) m_1 and m_2 are dependent on the same verb or noun, allowing for intervening nodes via modal, auxiliary, gerund and infinitive

[9] We changed the value of "full prev mention" from "numeric' to {yes, no, NA}.

constructions, (iii) m_1 is neither a child nor a parent of m_2, and (iv) m_1 occurs before m_2.

For Example 8 (slightly simplified) we extract the precedence relations shown in Table 5.

(8) She was sent by her mother to a white woman's house to do chores in exchange for meals and a place to sleep.

(She)$_{old} >_p$ (her mother)$_{med/synt}$
(She)$_{old} >_p$ (a white-woman's house)$_{new}$
(She)$_{old} >_p$ (chores)$_{new}$
(She)$_{old} >_p$ (exchange .....sleep)$_{new}$
(her mother)$_{med/synt} >_p$ (a white woman's house)$_{new}$
(chores)$_{new} >_p$ (exchange . . . sleep)$_{new}$
(meals)$_{new} >_p$ (a place to sleep)$_{new}$

Table 5: Precedence Relations for Example 8. *She* is a trace subject for *do*.

Proper names behave differently from common nouns. For example, they can occur at many different places in the clause when functioning as spatial or temporal scene-setting elements, such as *In New York*. We therefore exclude all precedence relations where one element of the pair is a proper name.

We extract 2855 precedence relations. Table 6 shows the statistics on precedence with the first mention in a pair in rows and the second in columns. Mediated and new mentions indeed rarely precede old mentions, so that precedence should improve separating of `old` vs other mentions.

	old	mediated	new
old	136	387	519
mediated	88	357	379
new	85	291	613

Table 6: Precedence relations in our corpus

5 Experiments

5.1 Experimental Setup

We use our gold standard corpus (see Section 3.3) via 10-fold cross-validation on documents for all experiments. Following Nissim (2006) and Rahman and Ng (2011), we perform all experiments on gold standard mentions and use the human WSJ syntactic annotation for feature extraction, when necessary. For the extraction of semantic class, we use OntoNotes entity type annotation for proper names and an automatic assignment of semantic class via WordNet hypernyms for common nouns.

Coarse-grained versions of all algorithms distinguish only between the three `old, mediated, new` categories. Fine-grained versions distinguish between the categories `old`, the six `mediated` subtypes, and `new`. We report overall accuracy as well as precision, recall and F-measure per category. Significance tests are conducted using McNemar's test on overall algorithm accuracy, at the level of 1%.

5.2 Local Classifiers

We reimplemented the algorithms in Nissim (2006) and Rahman and Ng (2011) as comparison baselines, using their feature and algorithm choices. Algorithm *Nissim* is therefore a decision tree J48 with standard settings in WEKA with the features in Table 4. Algorithm *RahmanNg* is an SVM with a composite kernel and one-vs-all training/testing (toolkit SVMLight). They use the features in Table 4 plus unigram and tree kernel features, described in Section 4.1. We add our additional set of *otherlocal* features to both baseline algorithms (yielding *Nissim+ol* and *RahmanNg+ol*) as they aim specifically at improving fine-grained classification.

5.3 Collective Classification

For incorporating our inter-mention links, we use a variant of Iterative Collective classification (ICA), which has shown good performance over a variety of tasks (Lu and Getoor, 2003) and has been used in NLP for example for opinion mining (Somasundaran et al., 2009). ICA is normally faster than Gibbs sampling and — in initial experiments — did not yield significantly different results from it.

ICA initializes each mention with its most likely IS, according to the local classifier and features. It then iterates a relational classifier, which uses both local and relational features (our *hasChild* and *precedes* features) taking IS assignments to neighbouring mentions into account. We use the *exist* aggregator to define the dependence between mentions.

We use NetKit (Macskassy and Provost, 2007) with its standard ICA settings for collective inference, as it allows direct comparison between local and collective classification. The relational classifiers are always exactly the same classifiers as the

		local						*collective*					
		Nissim			*Nissim+ol*			*Nissim+ol +hasChild*			*Nissim+ol +hasChild+precedes*		
		R	P	F	R	P	F	R	P	F	R	P	F
Coarse													
	old	82.2	86.4	84.2	81.2	88.6	84.8	81.7	88.6	**85.0**	80.9	89.1	84.8
	mediated	51.9	60.2	55.7	57.8	64.6	61.0	68.4	77.4	**72.6**	68.8	76.9	**72.6**
	new	74.2	63.6	68.5	78.4	67.3	72.4	87.7	75.1	**80.9**	87.9	75.0	**80.9**
	acc		69.0			72.3			**79.4**			**79.4**	
Fine													
	old	84.0	83.3	83.6	85.0	83.9	84.5	84.3	84.7	84.5	84.1	85.2	**84.6**
	med/knowledge	61.3	60.0	60.6	61.0	69.5	65.0	62.3	70.0	**65.9**	60.6	70.0	65.0
	med/synt	37.2	59.7	45.8	44.7	60.0	51.3	76.8	81.4	**79.0**	75.7	80.1	77.9
	med/agg	26.0	42.0	32.2	20.4	38.4	26.6	42.6	55.9	48.4	43.1	55.8	**48.7**
	med/func	0.0	NA	NA	32.3	65.6	43.3	33.8	53.7	41.5	35.4	53.5	**48.7**
	med/comp	0.4	7.70	0.7	79.0	82.6	80.0	80.6	82.9	**81.8**	81.4	82.0	81.7
	med/bridging	6.6	26.2	10.6	8.9	30.9	13.8	9.6	34.4	15.1	12.2	41.7	**18.9**
	new	82.6	61.0	70.2	82.7	65.1	72.8	88.0	74.0	**80.4**	87.7	73.3	79.8
	acc		66.6			70.0			**77.0**			76.8	

Table 7: Collective classification compared to Nissim's local classifier. Best performing algorithms are bolded.

local ones with the relational features added: thus, if the local classifier is a tree kernel SVM so is the relational one. One problem when using the SVM Tree kernel as relational classifier is that it allows only for binary classification so that we need to train several binary networks in a one-vs-all paradigm (see also (Rahman and Ng, 2011)), which will not be able to use the multiclass dependencies of the relational features to optimum effect.

5.4 Results

Table 7 shows the comparison of collective classification to local classification, using Nissim's framework and features, and Table 8 the equivalent table for Rahman and Ng's approach.

The improvements using the additional local features over the original local classifiers are statistically significant in all cases. In particular, the inclusion of semantic classes improves `mediated/knowledge` and `mediated/func`, and comparative anaphora are recognised highly reliably via a small set of comparative markers.

The *hasChild* relation leads to significant improvement in accuracy over local classification in all cases, showing the value of collective classification. The improvement here is centered on the categories of `mediated/synt` (for both cases) and `mediated/aggregate` (for *Nissim+ol+hasChild*) as well as their distinction from `new`.[10] It is also interesting that collective classification with a concise feature set and a simple decision tree as used in *Nissim+ol+hasChild*, performs equally well as *RahmanNg+ol+hasChild*, which uses thousands of unigram and tree features and a more sophisticated local classifier. It also shows more consistent improvements over all fine-grained classes.

The *precedes* relation does not lead to any further improvement. We investigated several variations of the precedence link, such as restricting it to certain grammatical relations, taking into account definiteness or NP type but none of them led to any improvement. We think there are two reasons for this lack of success. First, the precedence of `mediated` vs. `new` mentions does not follow a clear order and is therefore not a very predictive feature (see Table 6). At first, this seems to contradict studies such as Cahill and Riester (2009) that find a variety of precedences according to information status. However, many of the clearest precedences they find are more specific variants of the `old` $>_p$ `mediated` or `old` $>_p$ `new` precedence or they are preferences at an even finer level than the one we annotate, including for example the identification of generics. Second, the clear `old` $>_p$ `mediated`

[10] For *RhamanNg+ol+hasChild*, the `aggregate` class suffers from collective classification. We hypothesise that this is an artefact of the one-vs-all training/testing for rare categories.

		local						*collective*					
		RahmanNg			*RahmanNg+ol*			*RahmanNg+ol +hasChild*			*RahmanNg+ol +hasChild+precedes*		
		R	P	F	R	P	F	R	P	F	R	P	F
Coarse													
	`old`	81.3	90.1	85.5	82.6	91.4	**86.8**	83.5	87.8	85.6	82.9	87.2	85.0
	`mediated`	61.4	68.6	64.8	61.5	71.9	66.3	66.7	79.5	**72.6**	64.8	76.7	70.3
	`new`	82.1	69.9	75.5	84.9	70.1	76.8	89.0	74.9	**81.3**	86.9	73.5	79.6
	acc		74.9			76.3			**79.8**			78.3	
Fine													
	`old`	85.1	87.0	86.0	85.6	87.9	**86.7**	85.3	87.4	86.3	85.8	87.5	86.4
	`med/knowledge`	65.8	67.2	66.5	64.8	72.6	68.5	67.1	69.6	68.3	64.7	73.2	**68.7**
	`med/synt`	55.8	72.1	62.9	55.8	72.6	63.1	79.8	78.1	**78.9**	79.8	78.1	**78.9**
	`med/agg`	29.9	75.9	**42.9**	29.9	75.9	**42.9**	17.1	53.7	25.9	14.2	49.2	22.1
	`med/func`	27.7	38.3	32.1	38.5	69.4	**49.5**	40.0	44.1	42.0	40.0	40.0	40.0
	`med/comp`	25.3	86.5	39.1	76.7	82.2	**79.3**	74.3	62.7	68.0	74.3	62.7	68.0
	`med/bridging`	10.6	44.6	**17.1**	9.0	47.2	15.2	1.0	15.2	2.0	1.0	13.7	1.9
	`new`	87.3	66.3	75.4	89.0	67.8	77.0	89.2	74.6	**81.2**	89.2	74.6	**81.2**
	acc		72.6			74.6			**77.5**			77.4	

Table 8: Collective classification compared to Rahman and Ng's local classifier. Best performing algorithms are bolded.

and `old` $>_p$ `new` preferences are partially already captured by the local features, especially the grammatical role, as, for example, subjects are often both old as well as early on in a sentence.

With regard to fine-grained classification, many categories including comparative anaphora, are identified quite reliably, especially in the multiclass classification setting (*Nissim+ol+hasChild*). Bridging seems to be the by far most difficult category to identify with final best F-measures still very low. Most bridging mentions do not have any clear internal structure or external syntactic contexts that signal their presence. Instead, they rely more on lexical and world knowledge for recognition. Unigrams could potentially encapsulate some of this lexical knowledge but — without generalization — are too sparse for a relatively rare category such as bridging (6% of all mentions) to perform well. The difficulty of bridging recognition is an important insight of this paper as it casts doubt on the strategy in previous research to concentrate almost exclusively on antecedent selection (see Section 2).

6 Conclusions

We presented a new approach to information status classification in written text, for which we also provide the first reliably annotated English language corpus. Based on linguistic intuition, we define features for classifying mentions collectively. We show that our collective classification approach outperforms the state-of-the-art in coarse-grained IS classification by about 10% (Nissim, 2006) and 5% (Rahman and Ng, 2011) accuracy. The gain is almost entirely due to improvements in distinguishing between `new` and `mediated` mentions. For the latter, we also report the – to our knowledge – first fine-grained IS classification results.

Since the work reported in this paper relied – following Nissim (2006) and Rahman and Ng (2011) – on gold standard mentions and syntactic annotations, we plan to perform experiments with predicted mentions as well. We also have to improve the recognition of bridging, ideally combining recognition and antecedent selection for a complete resolution component. In addition, we plan to integrate IS resolution with our coreference resolution system (Cai et al., 2011) to provide us with a more comprehensive discourse processing system.

Acknowledgements. Katja Markert received a Fellowship for Experienced Researchers by the Alexander-von-Humboldt Foundation and Yufang Hou is funded by a PhD scholarship from the Research Training Group *Coherence in Language Processing* at Heidelberg University. We thank the Heidelberg Institute for Theoretical Studies for hosting Katja Markert and funding the annotation study, and the annotators for their diligent work.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Betty J. Birner and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. John Benjamins, Amsterdam, The Netherlands.

Aoife Cahill and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pages 817–825.

Jie Cai, Éva Mújdricza-Maydt, and Michael Strube. 2011. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 56–60.

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

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14,* Vancouver, B.C., Canada, 3–8 December, 2001, pages 625–632, Cambridge, Mass. MIT Press.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pages 236–243.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pages 320–327.

Claire Gardent and Hélène Manuélian. 2005. Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.

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

Iorn Korzen and Matthias Buch-Kromann. 2011. Anaphoric relations in the Copenhagen dependency treebanks. In S. Dipper and H. Zinsmeister, editors, *Corpus-based Investigations of Pragmatic and Discourse Phenomena*, volume 3 of *Bochumer Linguistische Arbeitsberichte*, pages 83–98. University of Bochum, Bochum, Germany.

Ivana Kruijff-Korbayová and Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information. Special Issue on Discource and Information Structure*, 12(3):149–259.

Knud Lambrecht. 1994. *Information Structure and Sentence Form*. Cambridge, U.K.: Cambridge University Press.

Qing Lu and Lise Getoor. 2003. Link-based classification. In *Proceedings of the 20th International Conference on Machine Learning,* Washington, D.C., 21–24 August 2003, pages 496–503.

Sofus A. Macskassy and Foster Provost. 2007. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.

Katja Markert and Malvina Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–401.

Josef Meyer and Robert Dale. 2002. Mining a corpus to support associative anaphora resolution. In *Proceedings of the 4th International Conference on Discourse Anaphora and Anaphor Resolution,* Lisbon, Portugal, 18–20 September, 2002.

Natalia M. Modjeska, Katja Markert, and Malvina Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* Sapporo, Japan, 11–12 July 2003, pages 176–183.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pages 9–16.

Vincent Ng. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 575–583.

Malvina Nissim, Shipara 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,* Lisbon, Portugal, 26–28 May 2004, pages 1023–1026.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July 2006, pages 94–012.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 272–279.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 143–150.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue,* Cambridge, Mass., 30 April – 1 May 2004, pages 154–162.

Scott Prevost. 1996. An information structural approach to spoken language generation. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* Santa Cruz, Cal., 24–27 June 1996, pages 294–301.

Ellen F. Prince. 1981. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York, N.Y.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In W.C. Mann and S.A. Thompson, editors, *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. John Benjamins, Amsterdam.

Altaf Rahman and Vincent Ng. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011, pages 1069–1080.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* La Valetta, Malta, 17–23 May 2010, pages 717–722.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco, 26 May – 1 June 2008, pages 2137–2142.

Ryohei Sasano and Sadao Kurohashi. 2009. A probabilistic model for associative anaphora resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 1455–1464.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009.

Ben Taskar, Pieter Abbeel, and Daphne Koller. 2002. Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence,* Edmonton, Alberta, Canada, 1-4 August 2002, pages 485–492.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.

Yiming Yang, Seán Slattery, and Rayid Ghani. 2002. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2-3):219–241.

Guodong Zhou and Fang Kong. 2009. Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 978–986.