Proceedings of the 13th Conference of the European Chapter of the
Association for Computational Linguistics (EACL), Avignon, France, April 2012, pp. 798--807.

Learning the Fine-Grained Information Status of Discourse Entities

Altaf Rahman and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{altaf,vince}@hlt.utdallas.edu

Abstract

While information status (IS) plays a crucial role in discourse processing, there have only been a handful of attempts to automatically determine the IS of discourse entities. We examine a related but more challenging task, *fine-grained* IS determination, which involves classifying a discourse entity as one of 16 IS *subtypes*. We investigate the use of rich knowledge sources for this task in combination with a rule-based approach and a learning-based approach. In experiments with a set of Switchboard dialogues, the learning-based approach achieves an accuracy of 78.7%, outperforming the rule-based approach by 21.3%.

1 Introduction

A linguistic notion central to discourse processing is *information status* (IS). It describes the extent to which a discourse entity, which is typically referred to by noun phrases (NPs) in a dialogue, is *available* to the hearer. Different definitions of IS have been proposed over the years. In this paper, we adopt Nissim et al.'s (2004) proposal, since it is primarily built upon Prince's (1992) and Eckert and Strube's (2001) well-known definitions, and is empirically shown by Nissim et al. to yield an annotation scheme for IS in dialogue that has good reproducibility.[1]

Specifically, Nissim et al. (2004) adopt a three-way classification scheme for IS, defining a discourse entity as (1) old to the hearer if it is known to the hearer and has previously been referred to in the dialogue; (2) new if it is unknown to her and has not been previously referred to; and (3) mediated (henceforth med) if it is newly mentioned in the dialogue but she can infer its identity from a previously-mentioned entity. To capture finer-grained distinctions for IS, Nissim et al. allow an old or med entity to have a *subtype*, which *subcategorizes* an old or med entity. For instance, a med entity has the subtype set if the NP that refers to it is in a set-subset relation with its antecedent.

IS plays a crucial role in discourse processing: it provides an indication of how a discourse model should be updated as a dialogue is processed incrementally. Its importance can be reflected in part in the amount of attention it has received in theoretical linguistics over the years (e.g., Halliday (1976), Prince (1981), Hajičová (1984), Vallduví (1992), Steedman (2000)), and in part in the benefits it can potentially bring to NLP applications. One task that could benefit from knowledge of IS is identity coreference: since new entities by definition have not been previously referred to, an NP marked as new does not need to be resolved, thereby improving the precision of a coreference resolver. Knowledge of *fine-grained* or *subcategorized* IS is valuable for other NLP tasks. For instance, an NP marked as set signifies that it is in a set-subset relation with its antecedent, thereby providing important clues for bridging anaphora resolution (e.g., Gasperin and Briscoe (2008)).

Despite the potential usefulness of IS in NLP tasks, there has been little work on *learning* the IS of discourse entities. To investigate the plausibility of learning IS, Nissim et al. (2004) annotate a set of Switchboard dialogues with such information[2], and subsequently present a

[1] It is worth noting that several IS annotation schemes have been proposed more recently. See Götze et al. (2007) and Riester et al. (2010) for details.

[2] These and other linguistic annotations on the Switchboard dialogues were later released by the LDC as part of the NXT corpus, which is described in Calhoun et al. (2010).

rule-based approach and a learning-based approach to acquiring such knowledge (Nissim, 2006). More recently, we have improved Nissim's learning-based approach by augmenting her feature set, which comprises seven string-matching and grammatical features, with lexical and syntactic features (Rahman and Ng, 2011; henceforth R&N). Despite the improvements, the performance on new entities remains poor: an F-score of 46.5% was achieved.

Our goal in this paper is to investigate *fine-grained IS determination*, the task of classifying a discourse entity as one of the 16 IS subtypes defined by Nissim et al. (2004).[3] Owing in part to the increase in the number of categories, fine-grained IS determination is arguably a more challenging task than the 3-class IS determination task that Nissim and R&N investigated. To our knowledge, this is the first empirical investigation of automated fine-grained IS determination.

We propose a *knowledge-rich* approach to fine-grained IS determination. Our proposal is motivated in part by Nissim's and R&N's poor performance on new entities, which we hypothesize can be attributed to their sole reliance on shallow knowledge sources. In light of this hypothesis, our approach employs *semantic* and *world* knowledge extracted from manually and automatically constructed knowledge bases, as well as *coreference* information. The relevance of coreference to IS determination can be seen from the definition of IS: a new entity is not coreferential with any previously-mentioned entity, whereas an old entity may. While our use of coreference information for IS determination and our earlier claim that IS annotation would be useful for coreference resolution may seem to have created a chicken-and-egg problem, they do not: since coreference resolution and IS determination can benefit from each other, it may be possible to formulate an approach where the two tasks can mutually bootstrap.

We investigate rule-based and learning-based approaches to fine-grained IS determination. In the rule-based approach, we manually compose rules to combine the aforementioned knowledge sources. While we could employ the same knowledge sources in the learning-based approach, we chose to encode, among other knowledge sources, the hand-written rules and their predictions directly as features for the learner. In an evaluation on 147 Switchboard dialogues, our learning-based approach to fine-grained IS determination achieves an accuracy of 78.7%, substantially outperforming the rule-based approach by 21.3%. Equally importantly, when employing these linguistically rich features to learn Nissim's 3-class IS determination task, the resulting classifier achieves an accuracy of 91.7%, surpassing the classifier trained on R&N's state-of-the-art feature set by 8.8% in absolute accuracy. Improvements on the new class are particularly substantial: its F-score rises from 46.7% to 87.2%.

2 IS Types and Subtypes: An Overview

In Nissim et al.'s (2004) IS classification scheme, an NP can be assigned one of three main types (old, med, new) and one of 16 subtypes. Below we will illustrate their definitions with examples, most of which are taken from Nissim (2003) or Nissim et al.'s (2004) dataset (see Section 3).

Old. An NP is marked is old if (i) it is coreferential with an entity introduced earlier, (ii) it is a generic pronoun, or (iii) it is a personal pronoun referring to the dialogue participants. Six subtypes are defined for old entities: identity, event, general, generic, ident_generic, and relative. In Example 1, *my* is marked as old with subtype identity, since it is coreferent with *I*.

(1) I was angry that he destroyed **my** tent.

However, if the markable has a verb phrase (VP) rather than an NP as its antecedent, it will be marked as old/event, as can be seen in Example 2, where the antecedent of *That* is the VP *put my phone number on the form*.

(2) They ask me to put my phone number on the form. **That** I think is not needed.

Other NPs marked as old include (i) relative pronouns, which have the subtype relative; (ii) personal pronouns referring to the dialogue participants, which have the subtype general, and (iii) generic pronouns, which have the subtype generic. The pronoun *you* in Example 3 is an instance of a generic pronoun.

(3) I think to correct the judicial system, **you** have to get the lawyer out of it.

Note, however, that in a coreference chain of generic pronouns, every element of the chain is

[3] One of these 16 classes is the new type, for which no subtype is defined. For ease of exposition, we will refer to the new type as one of the 16 subtypes to be predicted.

assigned the subtype ident_generic instead.

Mediated. An NP is marked as med if the entity it refers to has not been previously introduced in the dialogue, but can be inferred from already-mentioned entities or is generally known to the hearer. Nine subtypes are available for med entities: general, bound, part, situation, event, set, poss, func_value, and aggregation.

General is assigned to med entities that are generally known, such as *the Earth*, *China*, and most proper names. Bound is reserved for bound pronouns, an instance of which is shown in Example 4, where *its* is bound to the variable of the universally quantified NP, *Every cat*.

(4) Every cat ate **its** dinner.

Poss is assigned to NPs involved in intra-phrasal possessive relations, including prenominal genitives (i.e., X's Y) and postnominal genitives (i.e., Y of X). Specifically, Y will be marked as poss if X is old or med; otherwise, Y will be new. For example, in cases like *a friend's boat* where *a friend* is new, *boat* is marked as new.

Four subtypes, namely part, situation, event, and set, are used to identify instances of bridging (i.e., entities that are inferrable from a related entity mentioned earlier in the dialogue). As an example, consider the following sentences:

(5a) He passed by the door of Jan's house and saw that **the door** was painted red.

(5b) He passed by Jan's house and saw that **the door** was painted red.

In Example 5a, by the time the hearer processes the second occurrence of *the door*, she has already had a mental entity corresponding to *the door* (after processing the first occurrence). As a result, the second occurrence of *the door* refers to an old entity. In Example 5b, on the other hand, the hearer is not assumed to have any mental representation of the door in question, but she can infer that the door she saw was part of Jan's house. Hence, this occurrence of *the door* should be marked as med with subtype part, as it is involved in a part-whole relation with its antecedent.

If an NP is involved in a set-subset relation with its antecedent, it inherits the med subtype set. This applies to the NP *the house payment* in Example 6, whose antecedent is *our monthly budget*.

(6) What we try to do to stick to our monthly budget is we pretty much have **the house payment**.

If an NP is part of a situation set up by a previously-mentioned entity, it is assigned the subtype situation, as exemplified by the NP *a few horses* in the sentence below, which is involved in the situation set up by *John's ranch*.

(7) Mary went to John's ranch and saw that there were only **a few horses**.

Similar to old entities, an NP marked as *med* may be related to a previously mentioned VP. In this case, the NP will receive the subtype event, as exemplified by the NP *the bus* in the sentence below, which is triggered by the VP *traveling in Miami*.

(8) We were traveling in Miami, and **the bus** was very full.

If an NP refers to a value of a previously mentioned function, such as the NP *30 degrees* in Example 9, which is related to *the temperature*, then it is assigned the subtype func_value.

(9) The temperature rose to **30 degrees**.

Finally, the subtype aggregation is assigned to coordinated NPs if at least one of the NPs involved is not new. However, if all NPs in the coordinated phrase are new, the phrase should be marked as new. For instance, the NP *My son and I* in Example 10 should be marked as med/aggregation.

(10) I have a son ... **My son and I** like to play chess after dinner.

New. An entity is new if it has not been introduced in the dialogue and the hearer cannot infer it from previously mentioned entities. No subtype is defined for new entities.

There are cases where more than one IS value is appropriate for a given NP. For instance, given two occurrences of *China* in a dialogue, the second occurrence can be labeled as old/identity (because it is coreferential with an earlier NP) or med/general (because it is a generally known entity). To break ties, Nissim (2003) define a precedence relation on the IS subtypes, which yields a total ordering on the subtypes. Since all the old subtypes are ordered before their med counterparts in this relation, the second occurrence of *China* in our example will be labeled as old/identity. Owing to space limitations, we refer the reader to Nissim (2003) for details.

3 Dataset

We employ Nissim et al.'s (2004) dataset, which comprises 147 Switchboard dialogues. We parti-

tion them into a training set (117 dialogues) and a test set (30 dialogues). A total of 58,835 NPs are annotated with IS types and subtypes.[4] The distributions of NPs over the IS subtypes in the training set and the test set are shown in Table 1.

	Train	(%)	Test	(%)
old/identity	10236	(20.1)	1258	(15.8)
old/event	1943	(3.8)	290	(3.6)
old/general	8216	(16.2)	1129	(14.2)
old/generic	2432	(4.8)	427	(5.4)
old/ident_generic	1730	(3.4)	404	(5.1)
old/relative	1241	(2.4)	193	(2.4)
med/general	2640	(5.2)	325	(4.1)
med/bound	529	(1.0)	74	(0.9)
med/part	885	(1.7)	120	(1.5)
med/situation	1109	(2.2)	244	(3.1)
med/event	351	(0.7)	67	(0.8)
med/set	10282	(20.2)	1771	(22.3)
med/poss	1318	(2.6)	220	(2.8)
med/func_value	224	(0.4)	31	(0.4)
med/aggregation	580	(1.1)	117	(1.5)
new	7158	(14.1)	1293	(16.2)
total	50874	(100)	7961	(100)

Table 1: Distributions of NPs over IS subtypes. The corresponding percentages are parenthesized.

4 Rule-Based Approach

In this section, we describe our rule-based approach to fine-grained IS determination, where we manually design rules for assigning IS subtypes to NPs based on the subtype definitions in Section 2, Nissim's (2003) IS annotation guidelines, and our inspection of the IS annotations in the training set. The motivations behind having a rule-based approach are two-fold. First, it can serve as a baseline for fine-grained IS determination. Second, it can provide insight into how the available knowledge sources can be combined into prediction rules, which can potentially serve as "sophisticated" features for a learning-based approach.

As shown in Table 2, our ruleset is composed of 18 rules, which should be applied to an NP in the order in which they are listed. Rules 1–7 handle the assignment of old subtypes to NPs. For instance, Rule 1 identifies instances of old/general, which comprises the personal pronouns referring to the dialogue participants. Note that this and several other rules rely on coreference information, which we obtain from two sources: (1) chains generated automatically using the Stanford Deterministic Coreference Resolution System (Lee et al., 2011)[5], and (2) manually identified coreference chains taken directly from the annotated Switchboard dialogues. Reporting results using these two ways of obtaining chains facilitates the comparison of the IS determination results that we can realistically obtain using existing coreference technologies against those that we could obtain if we further improved existing coreference resolvers. Note that both sources provide *identity* coreference chains. Specifically, the gold chains were annotated for NPs belonging to old/identity and old/ident_generic. Hence, these chains can be used to distinguish between old/general NPs and old/ident_generic NPs, because the former are *not* part of a chain whereas the latter are. However, they cannot be used to distinguish between old/general entities and old/generic entities, since neither of them belongs to any chains. As a result, when gold chains are used, Rule 1 will classify all occurrences of "you" that are not part of a chain as old/general, regardless of whether the pronoun is generic. While the gold chains alone can distinguish old/general and old/ident_generic NPs, the Stanford chains cannot distinguish any of the old subtypes in the absence of other knowledge sources, since it generates chains for *all* old NPs regardless of their subtypes. This implies that Rule 1 and several other rules are only a very crude approximation of the definition of the corresponding IS subtypes.

The rules for the remaining old subtypes can be interpreted similarly. A few points deserve mention. First, many rules depend on the string of the NP under consideration (e.g., "they" in Rule 2 and "whatever" in Rule 4). The decision of which strings are chosen is based primarily on our inspection of the training data. Hence, these rules are partly data-driven. Second, these rules should be applied in the order in which they are shown. For instance, though not explicitly stated, Rule 3 is only applicable to the non-anaphoric "you" and "they" pronouns, since Rule 2 has already covered their anaphoric counterparts. Finally, Rule 7 uses non-anaphoricity as a test of old/event NPs. The

[4] Not all NPs have an IS type/subtype. For instance, a pleonastic "it" does not refer to any real-world entity and therefore does not have any IS, and so are nouns such as "course" in "of course", "accident" in "by accident", etc.

[5] The Stanford resolver is available from `http://nlp.stanford.edu/software/corenlp.shtml`.

1. **if** the NP is "I" or "you" and it is not part of a coreference chain, **then**
   subtype := old/general
2. **if** the NP is "you" or "they" and it is anaphoric, **then**
   subtype := old/ident_generic
3. **if** the NP is "you" or "they", **then**
   subtype := old/generic
4. **if** the NP is "whatever" or an indefinite pronoun prefixed by "some" or "any" (e.g., "somebody"), **then**
   subtype := old/generic
5. **if** the NP is an anaphoric pronoun other than "that", or its string is identical to that of a preceding NP, **then**
   subtype := old/ident
6. **if** the NP is "that" and it is coreferential with the immediately preceding word, **then**
   subtype := old/relative
7. **if** the NP is "it", "this" or "that", and it is not anaphoric, **then**
   subtype := old/event
8. **if** the NP is pronominal and is not anaphoric, **then**
   subtype := med/bound
9. **if** the NP contains "and" or "or", **then**
   subtype := med/aggregation
10. **if** the NP is a multi-word phrase that (1) begins with "so much", "something", "somebody", "someone", "anything", "one", or "different", or (2) has "another", "anyone", "other", "such", "that", "of" or "type" as neither its first nor last word, or (3) its head noun is also the head noun of a preceding NP, **then**
    subtype := med/set
11. **if** the NP contains a word that is a hyponym of the word "value" in WordNet, **then**
    subtype := med/func_value
12. **if** the NP is involved in a part-whole relation with a preceding NP based on information extracted from ReVerb's output, **then**
    subtype := med/part
13. **if** the NP is of the form "X's Y" or "poss-pro Y", where X and Y are NPs and poss-pro is a possessive pronoun, **then**
    subtype := med/poss
14. **if** the NP fills an argument of a FrameNet frame set up by a preceding NP or verb, **then**
    subtype := med/situation
15. **if** the head of the NP and one of the preceding verbs in the same sentence share the same WordNet hypernym which is not in synsets that appear one of the top five levels of the noun/verb hierarchy, **then**
    subtype := med/event
16. **if** the NP is a named entity (NE) or starts with "the", **then**
    subtype := med/general
17. **if** the NP appears in the training set, **then**
    subtype := its most frequent IS subtype in the training set
18. subtype := new

Table 2: Hand-crafted rules for assigning IS subtypes to NPs.

reason is that these NPs have VP antecedents, but both the gold chains and the Stanford chains are computed over NPs only.

Rules 8–16 concern med subtypes. Apart from Rule 8 (med/bound), Rule 9 (med/aggregation), and Rule 11 (med/func_value), which are arguably crude approximations of the definitions of the corresponding subtypes, the med rules are more complicated than their old counterparts, in part because of their reliance on the extraction of sophisticated knowledge. Below we describe the extraction process and the motivation behind them.

Rule 10 concerns med/set. The words and phrases listed in the rule, which are derived manually from the training data, provide suggestive evidence that the NP under consideration is a subset or a specific portion of an entity or concept mentioned earlier in the dialogue. Examples include "another bedroom", "different color", "somebody else", "any place", "one of them", and "most other cities". Condition 3 of the rule, which checks whether the head noun of the NP has been mentioned previously, is a good test for identity coreference, but since all the old entities have suppos-

edly been identified by the preceding rules, it becomes a reasonable test for set-subset relations.

For convenience, we identify part-whole relations in Rule 12 based on the output produced by ReVerb (Fader et al., 2011), an open information extraction system.[6] The output contains, among other things, relation instances, each of which is represented as a triple, <A,rel,B>, where rel is a relation, and A and B are its arguments. To preprocess the output, we first identify all the triples that are instances of the part-whole relation using regular expressions. Next, we create clusters of relation arguments, such that each pair of arguments in a cluster has a part-whole relation. This is easy: since part-whole is a transitive relation (i.e., <A,$part$,B> and <B,$part$,C> implies <A,$part$,C>), we cluster the arguments by taking the transitive closure of these relation instances. Then, given an NP NP_i in the test set, we assign med/part to it if there is a preceding NP NP_j such that the two NPs are in the same argument cluster.

In Rule 14, we use FrameNet (Baker et al., 1998) to determine whether med/situation should be assigned to an NP, NP_i. Specifically, we check whether it fills an argument of a frame set up by a preceding NP, NP_j, or verb. To exemplify, let us assume that NP_j is "capital punishment". We search for "punishment" in FrameNet to access the appropriate frame, which in this case is "rewards and punishments". This frame contains a list of arguments together with examples. If NP_i is one of these arguments, we assign med/situation to NP_i, since it is involved in a situation (described by a frame) that is set up by a preceding NP/verb.

In Rule 15, we use WordNet (Fellbaum, 1998) to determine whether med/event should be assigned to an NP, NP_i, by checking whether NP_i is related to an event, which is typically described by a verb. Specifically, we use WordNet to check whether there exists a verb, v, preceding NP_i such that v and NP_i have the same hypernym. If so, we assign NP_i the subtype med/event. Note that we ensure that the hypernym they share does not appear in the top five levels of the WordNet noun and verb hierarchies, since we want them to be related via a concept that is not overly general.

Rule 16 identifies instances of med/general. The majority of its members are *generally-known* entities, whose identification is difficult as it requires world knowledge. Consequently, we apply this rule only after all other med rules are applied. As we can see, the rule assigns med/general to NPs that are named entities (NEs) and definite descriptions (specifically those NPs that start with "the"). The reason is simple. Most NEs are generally known. Definite descriptions are typically not new, so it seems reasonable to assign med/general to them given that the remaining (i.e., unlabeled) NPs are presumably either new and med/general.

Before Rule 18, which assigns an NP to the new class by default, we have a "memorization" rule that checks whether the NP under consideration appears in the training set (Rule 17). If so, we assign to it its most frequent subtype based on its occurrences in the training set. In essence, this heuristic rule can help classify some of the NPs that are somehow "missed" by the first 16 rules.

The ordering of these rules has a direct impact on performance of the ruleset, so a natural question is: what criteria did we use to order the rules? We order them in such a way that they respect the total ordering on the subtypes imposed by Nissim's (2003) preference relation (see Section 3), except that we give med/general a lower priority than Nissim due to the difficulty involved in identifying generally known entities, as noted above.

5 Learning-Based Approach

In this section, we describe our learning-based approach to fine-grained IS determination. Since we aim to automatically label an NP with its IS subtype, we create one training/test instance from each hand-annotated NP in the training/test set. Each instance is represented using five types of features, as described below.

Unigrams (119704). We create one binary feature for each unigram appearing in the training set. Its value indicates the presence or absence of the unigram in the NP under consideration.

Markables (209751). We create one binary feature for each markable (i.e., an NP having an IS subtype) appearing in the training set. Its value is 1 if and only if the markable has the same string as the NP under consideration.

Markable predictions (17). We create 17 binary features, 16 of which correspond to the 16 IS subtypes and the remaining one corresponds to a "dummy subtype". Specifically, if the NP un-

[6] We use ReVerb ClueWeb09 Extractions 1.1, which is available from `http://reverb.cs.washington.edu/reverb_clueweb_tuples-1.1.txt.gz`.

der consideration appears in the training set, we use Rule 17 in our hand-crafted ruleset to determine the IS subtype it is most frequently associated with in the training set, and then set the value of the feature corresponding to this IS subtype to 1. If the NP does not appear in the training set, we set the value of the dummy subtype feature to 1.

Rule conditions (17). As mentioned before, we can create features based on the hand-crafted rules in Section 4. To describe these features, let us introduce some notation. Let Rule i be denoted by $A_i \longrightarrow B_i$, where A_i is the condition that must be satisfied before the rule can be applied and B_i is the IS subtype predicted by the rule. We could create one binary feature from each A_i, and set its value to 1 if A_i is satisfied by the NP under consideration. These features, however, fail to capture a crucial aspect of the ruleset: the ordering of the rules. For instance, Rule i should be applied only if the conditions of the first $i-1$ rules are not satisfied by the NP, but such ordering is not encoded in these features. To address this problem, we capture rule ordering information by defining binary feature f_i as $\neg A_1 \wedge \neg A_2 \wedge \ldots \neg A_{i-1} \wedge A_i$, where $1 \leq i \leq 16$. In addition, we define a feature, f_{18}, for the default rule (Rule 18) in a similar fashion, but since it does not have any condition, we simply define f_{18} as $\neg A_1 \wedge \ldots \wedge \neg A_{16}$. The value of a feature in this feature group is 1 if and only if the NP under consideration satisfies the condition defined by the feature. Note that we did not create any features from Rule 17 here, since we have already generated "markables" and "markable prediction" features for it.

Rule predictions (17). None of the features f_i's defined above makes use of the predictions of our hand-crafted rules (i.e., the B_i's). To make use of these predictions, we define 17 binary features, one for each B_i, where $i = 1, \ldots, 16, 18$. Specifically, the value of the feature corresponding to B_i is 1 if and only if f_i is 1, where f_i is a "rule condition" feature as defined above.

Since IS subtype determination is a 16-class classification problem, we train a multi-class SVM classifier on the training instances using SVM^{multiclass} (Tsochantaridis et al., 2004), and use it to make predictions on the test instances.[7]

6 Evaluation

Next, we evaluate the rule-based approach and the learning-based approach to determining the IS subtype of each hand-annotated NP in the test set.

Classification results. Table 3 shows the results of the two approaches. Specifically, row 1 shows their accuracy, which is defined as the percentage of correctly classified instances. For each approach, we present results that are generated based on gold coreference chains as well as automatic chains computed by the Stanford resolver.

As we can see, the rule-based approach achieves accuracies of 66.0% (gold coreference) and 57.4% (Stanford coreference), whereas the learning-based approach achieves accuracies of 86.4% (gold) and 78.7% (Stanford). In other words, the gold coreference results are better than the Stanford coreference results, and the learning-based results are better than the rule-based results. While perhaps neither of these results are surprising, we are pleasantly surprised by the *extent* to which the learned classifier outperforms the hand-crafted rules: accuracies increase by 20.4% and 21.3% when gold coreference and Stanford coreference are used, respectively. In other words, machine learning has "transformed" a ruleset that achieves mediocre performance into a system that achieves relatively high performance.

These results also suggest that coreference plays a crucial role in IS subtype determination: accuracies could increase by up to 7.7–8.6% if we solely improved coreference resolution performance. This is perhaps not surprising: IS and coreference can mutually benefit from each other.

To gain additional insight into the task, we also show in rows 2–17 of Table 3 the performance on each of the 16 subtypes, expressed in terms of recall (R), precision (P), and F-score (F). A few points deserve mention. First, in comparison to the rule-based approach, the learning-based approach achieves considerably better performance on almost all classes. One that is of particular interest is the new class. As we can see in row 17, its F-score rises by about 30 points. These gains are accompanied by a simultaneous rise in recall and precision. In particular, recall increases by about 40 points. Now, recall from the introduc-

[7] For all the experiments involving SVM^{multiclass}, we set C, the regularization parameter, to 500,000, since preliminary experiments indicate that preferring generalization to overfitting (by setting C to a small value) tends to yield poorer classification performance. The remaining learning parameters are set to their default values.

		Rule-Based Approach						Learning-Based Approach					
		Gold Coreference			Stanford Coreference			Gold Coreference			Stanford Coreference		
1	Accuracy	66.0			57.4			**86.4**			78.7		
	IS Subtype	R	P	F	R	P	F	R	P	F	R	P	F
2	old/ident	77.5	78.2	77.8	66.1	52.7	58.7	82.8	85.2	**84.0**	75.8	64.2	69.5
3	old/event	98.6	50.4	66.7	71.3	43.2	53.8	98.3	87.9	**92.8**	2.4	31.8	4.5
4	old/general	81.9	82.7	82.3	72.3	83.6	77.6	97.7	93.7	**95.6**	87.8	92.7	90.2
5	old/generic	55.9	55.2	55.5	39.2	39.8	39.5	76.1	87.3	**81.3**	39.9	85.9	54.5
6	old/ident_generic	48.7	77.7	59.9	27.2	51.8	35.7	57.1	87.5	**69.1**	47.2	44.8	46.0
7	old/relative	55.0	69.2	61.3	55.1	63.4	59.0	98.0	63.0	**76.7**	99.0	37.5	54.4
8	med/general	29.9	19.8	23.8	29.5	19.6	23.6	91.2	87.7	**89.4**	84.0	72.2	77.7
9	med/bound	56.4	20.5	30.1	56.4	20.5	30.1	25.7	65.5	**36.9**	2.7	40.0	5.1
10	med/part	19.5	100.0	32.7	19.5	100.0	32.7	73.2	96.8	**83.3**	73.2	96.8	**83.3**
11	med/situation	28.7	100.0	44.6	28.7	100.0	44.6	68.4	95.4	79.7	68.0	97.7	**80.2**
12	med/event	10.5	100.0	18.9	10.5	100.0	18.9	46.3	100.0	**63.3**	46.3	100.0	**63.3**
13	med/set	82.9	61.8	70.8	78.0	59.4	67.4	90.4	87.8	**89.1**	88.4	86.0	87.2
14	med/poss	52.9	86.0	65.6	52.9	86.0	65.6	93.2	92.4	92.8	90.5	97.6	**93.9**
15	med/func_value	81.3	74.3	77.6	81.3	74.3	77.6	88.1	85.9	**87.0**	88.1	85.9	**87.0**
16	med/aggregation	57.4	44.0	49.9	57.4	43.6	49.6	85.2	72.9	78.6	83.8	93.9	**88.6**
17	new	50.4	65.7	57.0	50.3	65.1	56.7	90.3	84.6	**87.4**	90.4	83.6	86.9

Table 3: IS subtype accuracies and F-scores. In each row, the strongest result, as well as those that are statistically indistinguishable from it according to the paired t-test ($p < 0.05$), are boldfaced.

tion that previous attempts on 3-class IS determination by Nissim and R&N have achieved poor performance on the new class. We hypothesize that the use of shallow features in their approaches were responsible for the poor performance they observed, and that using our knowledge-rich feature set could improve its performance. We will test this hypothesis at the end of this section.

Other subtypes that are worth discussing are med/aggregation, med/func_value, and med/poss. Recall that the rules we designed for these classes were only crude approximations, or, perhaps more precisely, simplified versions of the definitions of the corresponding subtypes. For instance, to determine whether an NP belongs to med/aggregation, we simply look for occurrences of "and" and "or" (Rule 9), whereas its definition requires that not all of the NPs in the coordinated phrase are new. Despite the over-simplicity of these rules, machine learning has enabled the available features to be combined in such a way that high performance is achieved for these classes (see rows 14–16).

Also worth examining are those classes for which the hand-crafted rules rely on sophisticated knowledge sources. They include med/part, which relies on ReVerb; med/situation, which relies on FrameNet; and med/event, which relies on WordNet. As we can see from the rule-based results (rows 10–12), these knowledge sources have yielded rules that achieved perfect precision but low recall: 19.5% for part, 28.7% for situation, and 10.5 for event. Nevertheless, the learning algorithm has again discovered a profitable way to combine the available features, enabling the F-scores of these classes to increase by 35.1–50.6%.

While most classes are improved by machine learning, the same is not true for old/event and med/bound, whose F-scores are 4.5% (row 3) and 5.1% (row 9), respectively, when Stanford coreference is employed. This is perhaps not surprising. Recall that the multi-class SVM classifier was trained to maximize classification accuracy. Hence, if it encounters a class that is both difficult to learn *and* is under-represented, it may as well aim to achieve good performance on the easier-to-learn, well-represented classes at the expense of these hard-to-learn, under-represented classes.

Feature analysis. In an attempt to gain additional insight into the performance contribution of each of the five types of features used in the learning-based approach, we conduct feature ablation experiments. Results are shown in Table 4, where each row shows the accuracy of the classifier trained on all types of features except for the one shown in that row. For easy reference, the accuracy of the classifier trained on all types of features is shown in row 1 of the table. According to the paired t-test ($p < 0.05$), performance drops significantly whichever feature type is removed. This suggests that all five feature types are contributing positively to overall accuracy. Also, the *markables* features are the least important in the presence of other feature groups, whereas *mark-*

Feature Type	Gold Coref	Stanford Coref
All features	86.4	78.7
−rule predictions	77.5	70.0
−markable predictions	72.4	64.7
−rule conditions	81.1	71.0
−unigrams	74.4	58.6
−markables	83.2	75.5

Table 4: Accuracies of feature ablation experiments.

Feature Type	Gold Coref	Stanford Coref
rule predictions	49.1	45.2
markable predictions	39.7	39.7
rule conditions	58.1	28.9
unigrams	56.8	56.8
markables	10.4	10.4

Table 5: Accuracies of classifiers for each feature type.

able predictions and *unigrams* are the two most important feature groups.

To get a better idea of the utility of each feature type, we conduct another experiment in which we train five classifiers, each of which employs exactly one type of features. The accuracies of these classifiers are shown in Table 5. As we can see, the *markables* features have the smallest contribution, whereas *unigrams* have the largest contribution. Somewhat interesting are the results of the classifiers trained on the rule conditions: the rules are far more effective when gold coreference is used. This can be attributed to the fact that the design of the rules was based in part on the definitions of the subtypes, which assume the availability of perfect coreference information.

Knowledge source analysis. To gain some insight into the extent to which a knowledge source or a rule contributes to the overall performance of the rule-based approach, we conduct ablation experiments: in each experiment, we measure the performance of the ruleset after removing a particular rule or knowledge source from it. Specifically, rows 2–4 of Table 6 show the accuracies of the ruleset after removing the memorization rule (Rule 17), the rule that uses ReVerb's output (Rule 12), and the cue words used in Rules 4 and 10, respectively. For easy reference, the accuracy of the original ruleset is shown in row 1 of the table. According to the paired t-test ($p < 0.05$), performance drops significantly in all three ablation experiments. This suggests that the memorization rule, ReVerb, and the cue words all contribute positively to the accuracy of the ruleset.

Feature Type	Gold Coref	Stanford Coref
All rules	66.0	57.4
−memorization	62.6	52.0
−ReVerb	64.2	56.6
−cue words	63.8	54.0

Table 6: Accuracies of the simplified ruleset.

	R&N's Features			Our Features		
IS Type	R	P	F	R	P	F
old	93.5	95.8	94.6	93.8	96.4	95.1
med	89.3	71.2	79.2	93.3	86.0	89.5
new	34.6	71.7	46.7	82.4	72.7	87.2
Accuracy	82.9			91.7		

Table 7: Accuracies on IS types.

IS type results. We hypothesized earlier that the poor performance reported by Nissim and R&N on identifying new entities in their 3-class IS classification experiments (i.e., classifying an NP as old, med, or new) could be attributed to their sole reliance on lexico-syntactic features. To test this hypothesis, we (1) train a 3-class classifier using the five types of features we employed in our learning-based approach, computing the features based on the Stanford coreference chains; and (2) compare its results against those obtained via the lexico-syntactic approach in R&N on our test set. Results of these experiments, which are shown in Table 7, substantiate our hypothesis: when we replace R&N's features with ours, accuracy rises from 82.9% to 91.7%. These gains can be attributed to large improvements in identifying new and med entities, for which F-scores increase by about 40 points and 10 points, respectively.

7 Conclusions

We have examined the fine-grained IS determination task. Experiments on a set of Switchboard dialogues show that our learning-based approach, which uses features that include handcrafted rules and their predictions, outperforms its rule-based counterpart by more than 20%, achieving an overall accuracy of 78.7% when relying on automatically computed coreference information. In addition, we have achieved state-of-the-art results on the 3-class IS determination task, in part due to our reliance on richer knowledge sources in comparison to prior work. To our knowledge, there has been little work on automatic IS subtype determination. We hope that our work can stimulate further research on this task.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-0812261 and IIS-1147644.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

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.

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

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

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

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 257–264.

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. 2007. Information structure. In *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS)*. Potsdam: Universitätsverlag Potsdam.

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

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

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34.

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*, pages 1023–1026.

Malvina Nissim. 2003. Annotation scheme for information status in dialogue. Available from `http://www.stanford.edu/class/cs224u/guidelines-infostatus.pdf`.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 94–102.

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

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In *Discourse Description: Diverse Analysis of a Fund Raising Text*, pages 295–325. John Benjamins, Philadelphia/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*, pages 1069–1080.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 717–722.

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

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the 21st International Conference on Machine Learning*, pages 104–112.

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