

# Unrestricted Bridging Resolution

Yufang Hou

IBM Research Ireland

yhou@ie.ibm.com

Katja Markert

Heidelberg University

Department of Computational

Linguistics

markert@cl.uni-heidelberg.de

Michael Strube

Heidelberg Institute for Theoretical

Studies

michael.strube@h-its.org

*In contrast to identity anaphors, which indicate coreference between a noun phrase and its antecedent, bridging anaphors link to their antecedent(s) via lexico-semantic, frame, or encyclopedic relations. Bridging resolution involves recognizing bridging anaphors and finding links to antecedents. In contrast to most prior work, we tackle both problems. Our work also follows a more wide-ranging definition of bridging than most previous work and does not impose any restrictions on the type of bridging anaphora or relations between anaphor and antecedent.*

*We create a corpus (ISNotes) annotated for information status (IS), bridging being one of the IS subcategories. The annotations reach high reliability for all categories and marginal reliability for the bridging subcategory. We use a two-stage statistical global inference method for bridging resolution. Given all mentions in a document, the first stage, bridging anaphora recognition, recognizes bridging anaphors as a subtask of learning fine-grained IS. We use a cascading collective classification method where (i) collective classification allows us to investigate relations among several mentions and autocorrelation among IS classes and (ii) cascaded classification allows us to tackle class imbalance, important for minority classes such as bridging. We show that our method outperforms current methods both for IS recognition overall as well as for bridging, specifically. The second stage, bridging antecedent selection, finds the antecedents for all predicted bridging anaphors. We investigate the phenomenon of semantically or syntactically related bridging anaphors that share the same antecedent, a phenomenon we call sibling anaphors. We show that taking sibling anaphors into account in a joint inference*

---

Submission received: 15 January 2017; revised version received: 16 January 2018; accepted for publication: 1 March 2018.

doi:10.1162/COLLa\_00315

© 2018 Association for Computational Linguistics

Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

*model improves antecedent selection performance. In addition, we develop semantic and salience features for antecedent selection and suggest a novel method to build the candidate antecedent list for an anaphor, using the discourse scope of the anaphor. Our model outperforms previous work significantly.*

## 1. Introduction

An **anaphor** is an expression whose interpretation depends upon a previous expression in the discourse (the **antecedent**). Figure 1 shows an excerpt of a news article with three anaphoric references: “its” is a pronominal anaphor referring back to the antecedent “The business,” which itself refers back to “The Bakersfield Supermarket.” Both of these two anaphors refer to the same entity as their antecedents. Differently, the bridging anaphor “friends” does not refer to the same entity as its antecedent “its owner.” The phenomena illustrated in (1) and (2) have attracted a lot of interest under the heading of **coreference resolution** (Hobbs 1978; Hirschman and Chinchor 1997; Soon, Ng, and Lim 2001; Lee et al. 2013, 2017, *inter alia*). This article, however, focuses on the phenomenon illustrated in (3), known as **bridging** (Clark 1975) or **associative anaphora** (Hawkins 1978). Bridging anaphors are anaphoric noun phrases that are not coreferent but instead linked via associative relations to the antecedent.

**Bridging resolution** has to recognize bridging anaphors and find links to their antecedents. In Example (1), the bridging anaphors **The windows**, **The carpets** and **walls** can be felicitously used thanks to their part-of relation to their antecedent *the Polish center*.<sup>1</sup>

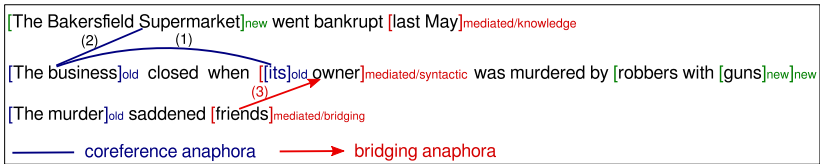
- (1) If Mr. McDonough’s plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw’s abundant rubble. [2 sent.] **The windows** will open. **The carpets** won’t be glued down and **walls** will be coated with non-toxic finishes.

Bridging plays an important role in establishing entity coherence in a text. Barzilay and Lapata (2008) model local coherence with the entity grid based on coreference only. However, Example (1) does not exhibit any coreferential entity coherence, and therefore entity coherence can only be established when bridging is resolved. Furthermore, text understanding applications such as textual entailment (Mirkin, Dagan, and Padó 2010), context question answering (Voorhees 2001), and opinion mining (Kobayashi, Inui, and Matsumoto 2007) have been shown to benefit from bridging resolution.

The main contributions presented in this article lie in the following aspects:

1. We present an English corpus (ISNotes) annotated for a wide range of information status (IS) categories as well as full anaphoric information for three anaphora types (coreference, bridging, and comparative; Section 3). Importantly, we impose no syntactic or relational restrictions on bridging—that is, bridging anaphora are not limited to definite noun phrases as in most previous work; antecedents can be noun phrases, verb

<sup>1</sup> All examples, if not specified otherwise, are from OntoNotes (Weischedel et al. 2011). Bridging anaphors are typed in **boldface**, antecedents in *italics* throughout this article.



**Figure 1**  
Coreference anaphora and bridging anaphora.

- phrases, or even clauses; and bridging relations are not restricted to meronymy. We show that bridging anaphora are very diverse. The overall annotation scheme is highly reliable, with the bridging category being annotated marginally reliably.<sup>2</sup> The corpus is available as an OntoNotes annotation layer via <http://www.h-its.org/en/research/nlp/isnotes-corpus/>.
2. We model **bridging anaphora recognition** as a subtask of learning fine-grained information status (Section 4). We integrate discourse structure, lexico-semantic, and genericity detection features into a **cascading collective classification algorithm**. Collective classification investigates relational autocorrelation among several IS classes whereas cascading classification addresses the multi-class imbalance problem, in particular the relative rarity of bridging compared to many other IS classes. Our model combines these two advantages by using binary classifiers for minority categories and a collective classifier for all categories. We beat current models both for overall IS classification accuracy as well as for bridging anaphora recognition on ISNotes.
  3. We explore a joint inference framework for **bridging antecedent selection** (Section 5). This model expresses an interesting topological property of bridging not used before—namely, that semantically or syntactically related anaphors are likely to share the same antecedent (such as **The windows** and **walls** in Example (1)). In addition, we develop semantic, syntactic, and salience features based on linguistic insights and present a novel method for constructing candidate antecedent lists, according to the anaphor’s **discourse scope**. Our model significantly outperforms prior work.
  4. Finally, we evaluate **bridging resolution** as a pipeline consisting of bridging recognition and antecedent selection (Section 6). This is the first full bridging resolution system that attempts the unrestricted phenomenon in a real setting.

All our experiments are performed on ISNotes and therefore all our claims hold only for the news genre. Although we believe the benefit of joint optimization to hold across other genres, several of our features are optimized for that particular corpus

<sup>2</sup> We consider annotation highly reliable when  $\kappa$  exceeds 0.80 and marginally reliable when between 0.67 and 0.80 (Carletta 1996). The interpretation of  $\kappa$  is still under discussion (Artstein and Poesio 2008).

and therefore our figures indicate the best possible performance of our approach. The adaptation to other corpora will likely need additional fine-tuning.<sup>3</sup>

*Connection to previous conference publications.* This article synthesizes Markert, Hou, and Strube (2012) and Hou et al. (2013a, 2013b). It provides more technical details, error analyses, and also includes the following new aspects. For the corpus, we now include a detailed analysis of our bridging cases (Section 3.3). In bridging recognition, we now use Markov Logic Networks instead of iterative collective classification to unify the approaches to the two tasks.<sup>4</sup> With regard to antecedent selection, we introduce several new features as well as the notion of using the **discourse scope** of an anaphor to adjust the set of potential antecedents it can refer back to (Section 5.3). We also now consider different evaluation paradigms dependent on whether one has access to full coreference information prior to bridging antecedent selection (mention-entity model) or not (mention-mention model), whereas before we only considered the mention-entity model (Section 5.4.4). Finally, we include a pipeline of the two models for bridging recognition and antecedent selection to evaluate performance of the full task (Section 6).<sup>5</sup>

## 2. Related Work

We first review theoretical studies related to bridging (Section 2.1) before discussing corpus studies in Section 2.2. Section 2.3 reviews automatic algorithms for bridging resolution and Section 2.4 discusses bridging and implicit semantic role labeling.

### 2.1 Bridging: Theoretical Studies

Theoretical studies on bridging include linguistic (Hawkins 1978; Prince 1981, 1992), psycholinguistic (Clark 1975; Clark and Haviland 1977; Garrod and Sanford 1982), pragmatic and cognitive (Erk  and Gundel 1987; Gundel, Hedberg, and Zacharski 2000; Matsui 2000; Schwarz 2000), as well as formal accounts (Hobbs et al. 1993; Bos, Buitelaar, and Mineur 1995; Asher and Lascarides 1998; L bner 1998; Cimiano 2006; Irmer 2009).

Our concept of bridging is closest to the notions of associative anaphora in Hawkins (1978) and (noncontained) inferrables in Prince (1981): noun phrases (NPs) that are not coreferent to a previous mention but the referent of which is identifiable via a lexico-semantic, frame, or encyclopedic relation to a previous mention, with this relation not being syntactically expressed.

Relation types used are very diverse and antecedents can be noun phrases, verb phrases, or even whole sentences (Clark 1975; Asher and Lascarides 1998, *inter alia*). Several studies, such as Hawkins (1978) and L bner (1998), limit bridging to definite NPs; we, however, believe that there is no clear difference in information status between **the windows**, on the one hand, and **walls**, on the other hand, in Example (1).<sup>6</sup>

3 Unfortunately, as we explain in Section 2.2, no other English corpus that is immediately usable for the full problem of bridging resolution is currently available for us to test our system on.

4 Quantitative results for bridging recognition are very similar to the previous framework, however.

5 We do not include the work that we conducted previously (Hou, Markert, and Strube 2014; Hou 2016), as these follow very different paradigms, using rule-based and neural network approaches, respectively.

None of these approaches outperform our work in this article.

6 Prince (1992) also gives examples of indefinite bridging cases, so our observation is not new.

Our bridging notion differs from Clark (1975) in that we do not include coreferential cases: We believe coreference is different both from an IS viewpoint (always being discourse-old) as well as from a computational perspective in that coreference needs different methods to resolve than bridging.

## 2.2 Bridging: Corpus Studies

Fraurud (1990) annotated first-mentioned NPs (which included bridging) versus subsequent mention NPs. Thirty-six percent of first-mentioned definite NPs have interpretations that “appear to involve a relation to contextual elements outside the definite NP itself” (Fraurud 1990, page 406), similar to our bridging definition.

The Vieira/Poesio data set (Poesio and Vieira 1998) contains 150 anaphoric definite NPs without a head match to their antecedents. These cases include what we call bridging as well as coreferential NPs without the same head. We will call this definition of bridging **lenient bridging** from now on. The corpus was used later to develop computational models (Section 2.3). In a second experiment, the authors delimited bridging proper from coreferential cases with very low agreement (31% per-class agreement).

Similarly, bridging recognition proved difficult for annotators of the GNOME corpus (Poesio 2004), where only 22% of bridging references were annotated in the same way by both annotators, although bridging relations were limited to set membership, subset, and generalized possession (part-of and ownership relations).

Nissim et al. (2004) is the first large-scale annotation study for IS for English. Based on Prince (1992) and Eckert and Strube (2000), they annotated NP types with three main categories: an **old** entity is known to the hearer and has been mentioned in the conversation; a **new** entity is unknown to the hearer and has not been previously referred to; a **mediated** entity is newly mentioned in the dialogue but is inferrable from previously mentioned entities, or generally known to the hearer. Four of the nine subtypes of the *mediated* category (*part*, *set*, *situation*, and *event*) include bridging instances. Nissim et al. (2004) reported high agreement for the overall fine-grained IS annotation (with  $\kappa = 0.788$ ) on 147 Switchboard dialogues (LDC 1993). The  $\kappa$  scores for the four bridging subtypes are mostly marginally reliable, between 0.594 and 0.794. However, the corpus cannot easily be used for a computational study of bridging anaphora resolution for the following reasons. First, antecedents for bridging NPs are not annotated. Second, the four subcategories used to mark up bridging also contain non-anaphoric cases, such as syntactically linked part-of relations (Example: *the house's door*). In addition, any such study would be limited with regard to relation types as several of the bridging cases are only annotated if the relation to the antecedent is part of certain knowledge bases (i.e., part-of relations must be part of WordNet and situation relations part of FrameNet).

The German DIRNDL corpus (Eckart, Riester, and Schweitzer 2012; Björkelund et al. 2014) contains IS annotations for all NPs following the scheme by Riester, Lorenz, and Seemann (2010). Bridging is one IS category but only used for definite expressions. They achieved a kappa score of 0.78 for six top-level categories. However, the confusion matrix in Riester, Lorenz, and Seemann (2010) shows that the anaphoric bridging category is frequently confused with other categories: The two annotators agreed on fewer than a third of bridging anaphors.

These previous corpus studies on bridging differ from ours in several ways. First, the definition of bridging is sometimes extended to include coreferential NPs

with lexical variety (Vieira 1998) or non-anaphoric NPs (Nissim et al. 2004). Second, they put more restrictions on bridging than we do, limiting to definite NP anaphora (Poesio and Vieira 1998; Gardent and Manuélian 2005; Caselli and Prodanof 2006; Riester, Lorenz, and Seemann 2010), to NP antecedents (all prior work), or to few relation types between anaphor and antecedent (Poesio 2004). Apart from these differences in definition of bridging, often reliability is not measured or low, especially for bridging recognition (Fraurud 1990; Poesio and Vieira 1998; Gardent and Manuélian 2005; Nedoluzhko, Mírovský, and Pajas 2009; Riester, Lorenz, and Seemann 2010).

### 2.3 Bridging: Computational Approaches

Most computational approaches for resolving bridging focus on antecedent selection. Some handle bridging anaphora recognition when recognizing fine-grained IS. Only a few works tackle full bridging resolution—that is, recognizing bridging anaphors and finding links to antecedents.

*Bridging anaphora recognition.* Fine-grained IS classification for Switchboard (Nissim et al. 2004) has been implemented via a combination of rules and a multiclass SVM (Rahman and Ng 2012). F-scores for the four categories that include bridging (*part, situation, event, set*) ranged from 63.3 to 87.2. These results do not necessarily reflect the real difficulty of the problem, however, because of the restrictions posed on bridging in the underlying annotation and the inclusion of non-anaphoric cases (Section 2.2).

Cahill and Riester (2012) trained a CRF model for fine-grained IS classification on the German DIRNDL radio news corpus (Riester, Lorenz, and Seemann 2010), making use of the assumption that IS classes within sentences tend to follow certain orderings, for example, *old* > *mediated* > *new*. They did not report the result for the *bridging* subcategory.

An attention-based long short-term memory model with pre-trained word embeddings and simple features achieved competitive results on ISNotes compared to our collective classification approach (Hou 2016).

*Bridging antecedent selection.* Based on the Vieira/Poesio data set (Section 2.2), various studies resolved “lenient” definite bridging references. Vieira and Teufel (1997) and Poesio, Vieira, and Teufel (1997) used heuristics for antecedent selection, exploiting WordNet relations such as synonymy/hyponymy/meronymy. Schulte im Walde (1998) used word clustering. The bridging anaphors were resolved to the closest antecedent candidate in a high-dimensional space, the best result being an accuracy of 22.7%.

Poesio et al. (2002) and Markert, Nissim, and Modjeska (2003) acquired mereological knowledge for bridging resolution by using syntactic patterns (such as *the NP of NP*) on the British National Corpus and the Web, respectively. All of this work was done on small data sets, numbering in the 10s for test bridging cases when excluding coreferential cases.

Another line of work applied machine learning techniques. The pairwise model in Poesio et al. (2004a) combines lexico-semantic and salience features to resolve mereological bridging in the GNOME corpus. However, their results came from a limited evaluation setting: In the first two experiments they distinguished only between the correct antecedent and *one* or *three* false candidates. The more realistic scenario of finding the correct antecedent among all possible candidates was tried for just six

bridging anaphors. On the basis of this method, Lassalle and Denis (2011) developed a system that resolves mereological bridging in French, with meronymic information extracted from raw texts using a bootstrapping method. They reported an accuracy of 23% for over 300 meronymy bridging anaphors using the realistic evaluation scenario.

*Full bridging resolution.* The rule-based system for processing definite NPs in Vieira and Poesio (2000) includes bridging cases (using the lenient definition of bridging discussed in the previous sections) but they do not report results for the bridging category.

Hahn, Strube, and Markert (1996) distinguish bridging resolution from other anaphora resolution. Their rule-based framework integrates language-independent conceptual criteria and language-dependent functional constraints. Their conceptual criteria were based on a knowledge base from the information technology domain that consists of 449 concepts and 334 relations. They focused on definite bridging anaphora and certain types of relations only (e.g., *has-property*, *has-physical-part*). On a small-scale technical domain data set (5 texts in German with 109 bridging anaphors), they achieved a recall of 55.0% and precision of 73.2%. Although the results seem satisfactory, the system is heavily dependent on the domain knowledge resource.

Sasano and Kurohashi (2009) resolved bridging and zero anaphora in Japanese simultaneously, using automatically acquired case frames in a probabilistic model. Although it is not clear how bridging anaphora are distributed in their corpus and whether this approach can be effectively applied to other languages, the lexical knowledge resource constructed is general and can capture diverse bridging relations.

Rösiger and Teufel (2014) extended a coreference resolution system with semantic features from WordNet (e.g., *hypernymy*, *meronymy*) to find bridging links in scientific text, considering definite NPs only. They used the CoNLL scorer for evaluation. However, a coreference resolution system and evaluation metric are not suitable for bridging resolution because bridging is not a set problem.

*Discussion.* Our study departs from related work by modeling bridging on the discourse level without limiting it to definite NPs or to certain bridging relations (e.g., *part-of*). For bridging anaphora recognition, our cascading collective classification model (Section 4) addresses multi-class imbalance while keeping the strength of collective classification. For bridging antecedent selection, our joint inference model (Section 5) integrates bridging resolution with clustering anaphors that share the same antecedent. Furthermore, unlike previous work that uses a sentence window to form the set of antecedent candidates, we propose a method to select antecedent candidates using a flexible notion of discourse scope of an anaphor. The latter makes use of the discourse relation Expansion and models salience.

## 2.4 Implicit Semantic Role Labeling

Semantic role labeling is the task of assigning semantic roles (such as *Agent* or *Theme*) to the semantic arguments associated with a predicate (e.g., a verb or a noun). In frame semantics (Baker, Fillmore, and Lowe 1998), core semantic roles (also called *Core Frame Elements*) are essential to the meaning of semantic situations while non-core semantic roles (e.g., *time*, *manner*) are less central.

The majority of work on semantic role labeling only recognizes semantic arguments from the sentence where the predicate is present and thus ignores arguments from the wider discourse context. Ruppenhofer et al. (2010) organized a shared task to

address the issue of non-local (implicit) argument identification for nominal and verbal predicates. There is partial overlap between bridging resolution and implicit semantic role labeling (i.e., in some bridging cases, antecedents are implicit semantic roles of bridging anaphors). However, bridging resolution considers all possible nominal bridging anaphors in running text. Some bridging anaphors are not considered “nominal predicates” in (implicit) semantic role labeling, for example, **One man** in Example (2).

- (2) Still, *employees* do occasionally try to smuggle out a gem or two. **One man** wrapped several diamonds in the knot of his tie.

In addition, implicit semantic role labeling for nominal predicates tries to link all possible implicit core roles for the nominal predicate in question. Yet not every nominal predicate under consideration is a bridging anaphor.

Despite differences between implicit semantic role labeling and bridging resolution, these two tasks can benefit from each other. We explore statistics from NomBank (Meyers et al. 2004) to predict bridging anaphors (Section 4.3.2). Some of our features for bridging antecedent selection are inspired by Laparra and Rigau (2013) (Section 5.2.2).

### 3. ISNotes: A Corpus for Information Status

ISNotes contains 50 texts from the Wall Street Journal portion of OntoNotes (Weischedel et al. 2011), in which all mentions (10,980 overall) are annotated for IS. The corpus can be downloaded from <http://www.h-its.org/en/research/nlp/isnotes-corpus/>.

#### 3.1 ISNotes Annotation Scheme

*Information status in ISNotes.* Information status describes the degree to which a discourse entity is available to the hearer regarding the speaker’s assumption about the hearer’s knowledge and beliefs (Prince 1992; Nissim et al. 2004). We distinguish eight IS categories, inspired by Nissim et al. (2004), although with some variations.

A mention is *old* if it is either coreferent with a previous mention (based on the OntoNotes coreference annotation), or if it is a generic or deictic pronoun.

Mediated mentions have not been mentioned before but are not autonomous—that is, they can only be correctly interpreted by reference to another mention or to prior world knowledge. ISNotes distinguishes six subcategories of mediated mentions:

- mediated/worldKnowledge (abbreviated as mediated/WK) mentions are generally known to the hearer. This category includes many proper names, such as *Poland*.
- mediated/syntactic mentions are syntactically linked via a possessive relation, a proper name premodification or a prepositional phrase postmodification to other old or mediated mentions, such as:
  - *[[their]<sub>old</sub> liquor store]<sub>mediated/syntactic</sub>*
  - *[the [Federal Reserve]<sub>mediated/WK</sub> boss]<sub>mediated/syntactic</sub>*, and
  - *[the main artery into [San Francisco]<sub>mediated/WK</sub>]<sub>mediated/syntactic</sub>*
- mediated/comparative mentions are non-coreference anaphors where the anaphor is compared to the antecedent (and where both are therefore often of the same semantic type). They usually include a premodifier or head



that makes clear that this entity is compared to a previous one, such as **others** in Example (3).<sup>7,8</sup>

- mediated/bridging mentions are non-coreference anaphors where a frame, lexico-semantic, or world knowledge relation holds between anaphor and antecedent, such as **the streets** in Example (4) and **The reason** in Example (5).
  - mediated/aggregate mentions are coordinated mentions where at least one element in the conjunction is old or mediated, such as *[Not only [George Bush]<sub>mediated/WK</sub> but also [Barack Obama]<sub>mediated/WK</sub>]<sub>mediated/aggregate</sub>*.
  - mediated/function mentions refer to a value of a previously mentioned function (e.g., *3 points* in Example (6)). The function needs to be able to rise and fall (e.g., *were down* in Example (6)).
- (3) As the death toll from last week's tremblor climbed to 61, the condition of *freeway survivor Buch Helm*, who spent four days trapped under rubble, improved, hospital officials said. Rescue crews, however, gave up hope that **others** would be found.
- (4) *Oranjemund, the mine headquarters*, is a lonely corporate oasis of 9,000 residents. Jackals roam **the streets** at night ...
- (5) The Bakersfield supermarket *went out of business* last May. **The reason** was not high interest rates or labor costs.
- (6) IBM shares were down 3 points.

New mentions have not yet been introduced in the discourse and the entity they refer to cannot be inferred from either previously mentioned entities/events or general world knowledge.

*Antecedents for mediated/bridging and mediated/comparative.* Antecedents for both mediated/bridging and mediated/comparative categories are annotated.<sup>9</sup> The antecedents can be NPs (Example (4)), verb phrases (e.g., *went out of business* in Example (5)), or even clauses. If an NP antecedent has several instantiations within the text, ISNotes chooses the one which is the closest to the bridging or comparative mention. Other instantiations can be inferred from the coreference annotation. Sometimes, several non-coreferent antecedents are annotated for a mediated/bridging mention when the antecedents fill core semantic roles. In Example (7), two antecedents are necessary to interpret the bridging anaphor **Domestic demand**.<sup>10</sup>

- (7) *Japan's production of cars, trucks and buses* in September fell 4.4% from a year ago. [...] **Domestic demand** continues to grow, but ...

<sup>7</sup> Comparative anaphors are typed in **boldface**, antecedents in *italics*.

<sup>8</sup> Nissim et al. (2004) view comparative anaphora as a subset of bridging. We distinguish them as their recognition (via lexical clues) and their resolution (often type matches) differ from other bridging cases.

<sup>9</sup> Antecedents for old mentions are from the OntoNotes coreference annotation.

<sup>10</sup> In ISNotes, only 2.6% of bridging anaphors have at least two antecedents. Our automatic system currently cannot deal with such cases—we leave this for future work.

**Table 1**  
Agreement results, overall (top) and for individual categories (bottom).

		A-B	A-C	B-C
Overall	Percentage coarse	87.5	86.3	86.5
	κ coarse	77.3	75.2	74.7
	Percentage fine	86.6	85.3	85.7
	κ fine	80.1	77.7	77.3
Individual Categories	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/worldKnowledge	82.1	78.4	74.1
	κ mediated/syntactic	88.4	87.8	87.6
	κ mediated/aggregate	87.0	85.4	86.0
	κ mediated/function	6.0	83.2	6.9
	κ mediated/comparative	81.8	78.3	81.2
	κ mediated/bridging	70.8	60.6	62.3

3.2 Agreement Study

An agreement study was carried out among three annotators. Annotator A is the scheme developer and a computational linguist. Annotators B and C have no linguistic training or education. Annotator A and B are fluent English speakers, living in English-speaking countries, but are not native speakers. Annotator C is a native speaker of English.

All potential mentions were pre-marked automatically using the WSJ syntactic noun phrase annotation. All non-initial mentions in an OntoNotes coreference chain were pre-marked as *old*. The annotation task consisted of excluding all non-mentions (such as non-referential *it*) and marking all mentions for their information status as well as the antecedents for comparative and bridging anaphora. The scheme was developed on nine texts, which were also used for training the annotators. Inter-annotator agreement was measured on 26 new texts, which included 5,905 potential mentions. The annotations of 1,499 of these were carried over from OntoNotes coreference annotation, leaving 4,406 potential mentions for annotation and agreement measurement.

Table 1 (top) shows percentage agreement as well as Cohen’s κ (Artstein and Poesio 2008) between all three possible annotator pairings at the coarse-grained (four categories: non-mention, old, new, mediated) and the fine-grained level (nine categories: non-mention, old, new and the six mediated subtypes). As our category distribution is highly unbalanced, Cohen’s kappa is necessary to report as it corrects for chance agreement achieved by just using majority categories.<sup>11</sup> Table 1 (bottom) shows individual category agreement, computed by merging all categories but one and then

<sup>11</sup> The κ values for the fine-grained scheme are higher than for the coarse-grained one. The hierarchical scheme is organized such that a category lower down the tree is more often confused with a category higher up in a different branch of the tree than with its direct siblings in the tree (i.e., mediated/bridging mentions are often confused with new mentions whereas some mediated categories such as mediated/syntactic or mediated/comparative are very easy to recognize).

**Table 2**  
IS distribution in ISNotes. The last column indicates the percentage of each IS category relative to the total number of mentions.

Texts	50	
Sentences	1,726	
Mentions	10,980	
old	3,237	29.5%
coreferent	3,143	28.6%
generic or deictic pronoun	94	0.9%
mediated	3,708	33.8%
syntactic	1,592	14.5%
world knowledge	924	8.4%
bridging	663	6.0%
comparative	253	2.3%
aggregate	211	1.9%
function	65	0.6%
new	4,035	36.7%

computing  $\kappa$  as usual. High reliability is achieved for most individual categories.<sup>12</sup> The reliability of the category bridging is marginally reliable and more annotator-dependent, although higher than other previous attempts at bridging annotation (Poesio 2003; Gardent and Manuélian 2005; Riester, Lorenz, and Seemann 2010). The agreement of selecting bridging antecedents is around 80% for all annotator pairings.

We investigated disagreements between Annotators A and B in bridging recognition: Almost all cases are instances where one annotator identified bridging and the other one new. Particularly frequent were borderline cases where the whole document had one major focus and subsequent NPs with a semantic relation to that focus could be seen either as new (interpretable without the major focus) or bridging. As an example, consider a document on the company *Toyota* and a later sentence stating *Output had gone down*. According to our guidelines, most of these cases are bridging, but they are easily overlooked.

The bridging annotations of the pairing A-B were used to create a consistent gold standard of the 35 texts (9 training, 26 testing), discussing all disagreed items between the annotators. Finally, Annotator A annotated a further 15 texts singly.

3.3 Corpus Analysis

*IS distribution.* Table 2 shows the class distribution. New mentions are the largest category (36.7%). Syntactic mentions are the largest mediated category.

<sup>12</sup> The low reliability of category function, when involving Annotator B, is explained by Annotator B forgetting about this category completely and only using it once. When two annotators remembered the category, it was easy to annotate reliably ( $\kappa$  83.2 for the pairing A-C).

*Bridging anaphora modification.* Bridging anaphors can be definite NPs (Examples (9) and (11)), indefinite NPs (Example (10)), or bare NPs (Examples (8), (12)–(14)). The only syntactic property shared is that bridging anaphora tend to have a simple internal structure with regard to modification. They are also easily confused with generics: *friends* is used as a bridging anaphor in Example (14) but generically in Example (15).

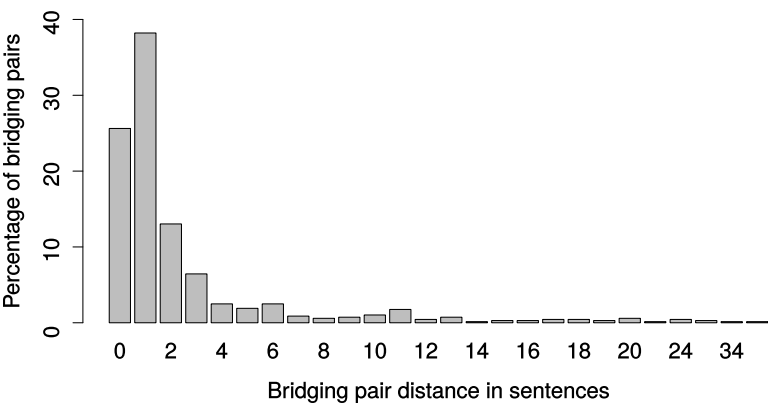
- (8) In June, farmers held onto *meat, milk and grain*, waiting for July’s usual state-directed price rises. The Communists froze **prices** instead.
- (9) To reduce it at *the fund’s building*, workers rubbed beeswax instead of polyurethane on the floors in the executive’s office. [1 sent.] **The budget** was only \$400,000.
- (10) Still, *employees* do occasionally try to smuggle out a gem or two. [2 sent.] **A food caterer** stashed stones in the false bottom of a milk pail.
- (11) *His truck* is parked across the field, in a row of grain sellers. [2 sent.] The farmer at **the next truck** shouts, “Wheat!”
- (12) The survey found that over a three-year period 22% of *the firms* said employees or owners had been robbed on their way to or from work or while on the job. [1 sent.] Crime was the reason that **26%** reported difficulty recruiting personnel and that **19%** said they were considering moving.
- (13) Mr. Leavitt, 37, was elected chairman earlier this year by *the company’s* new board [...] His father was **chairman** and **chief executive** until his death in an accident five years ago.
- (14) *She* made money, but spent more. **Friends** pitched in.
- (15) Friends are part of the glue that holds life and faith together.

Table 3 shows the bridging anaphora distribution with regard to determiners: Only 38.5% of bridging anaphors are modified by *the*, 44.9% of bridging anaphors are not modified by any determiners. This calls into question the strategy of several prior approaches (Vieira and Poesio 2000; Lassalle and Denis 2011; Cahill and Riester 2012) to limit themselves to bridging anaphors modified by *the*.

*Bridging pair distance.* We define the distance between a bridging anaphor and its antecedent as the distance between the anaphor and its closest preceding antecedent instantiation. The distribution of the distance for all 683 anaphor-antecedent pairs is

**Table 3**  
Bridging anaphora distribution with respect to determiners.

NP Type	Bridging Anaphors	
The	255	(38.5%)
A/An	70	(10.6%)
Other determiner	40	(6.0%)
No determiner	298	(44.9%)
Total	663	(100.0%)



**Figure 2**  
The distribution of anaphor-antecedent distances in sentences.

**Table 4**  
Bridging pair distribution with respect to relation types.

Relation Type	Bridging Pairs	
Action	16	(2.3%)
Set/Membership	45	(6.6%)
part-of/attribute-of	92	(13.5%)
Other	530	(77.6%)
Total	683	(100.0%)

shown in Figure 2.<sup>13</sup> We see that 77% of anaphors have antecedents occurring in the same or up to two sentences prior to the anaphor, although that still leaves a substantial number of instances that need relatively distant antecedents.

*Bridging relations.* The semantic relations between anaphor and antecedent are extremely diverse. Among 683 bridging pairs, only 2.3% correspond to an action, 6.6% to a set membership (see Example (2)) and 13.5% to a part-of/attribute-of relation between anaphor and antecedent (Table 4). A total of 77.6% of bridging relations fall under the category “other,” without further distinction. This includes encyclopedic relations such as *restaurant* – **the waiter** as well as context-specific relations such as *palms* – **the thieves**. Among all bridging antecedents, only 39 are represented by verbs or clauses.

*Sibling anaphors.* We call bridging anaphors “siblings” if they share the same antecedent (entity), and “non-siblings” are anaphors that do not share an antecedent with any other anaphor. In Example (1), **The windows**, **The carpets**, and **walls** are sibling anaphors.

13 A small portion of anaphors has more than one antecedent. Therefore, the number of anaphor-antecedent pairs (683) is slightly higher than the number of anaphors (663) (Section 3.1 and Example (7)).

In ISNotes, 61.4% of the bridging anaphors are siblings and we will use this to good effect in our model for bridging antecedent selection.

## 4. Information Status and Bridging Anaphora Recognition

For IS recognition, each mention is assigned one of the eight classes *old*, *mediated/syntactic*, *mediated/WK*, *mediated/bridging*, *mediated/comparative*, *mediated/aggregate*, *mediated/function*, and *new*. We make contributions to bridging recognition as well as for IS recognition in general.

### 4.1 Motivation for the Task

IS recognition can be beneficial for NLP tasks such as determining constituent order in generation (Cahill and Riester 2009) or coreference resolution (Rahman and Ng 2011). Treating bridging anaphora recognition as part of IS recognition prior to antecedent selection implies that it is possible to recognize bridging anaphors without knowing the antecedent. Predicting bridging anaphors and their antecedents jointly might be more attractive because some antecedents could trigger subsequent bridging. In Example (1), the antecedent *the Polish center* could trigger the anaphor **walls**. However, bridging anaphors can be solely indicated by referential patterns as nonsense Example (16) shows that **the wug** is clearly a bridging anaphor although we do not know the antecedent.<sup>14</sup>

(16) The blicket couldn't be connected to the dax. **The wug** failed.

In a similar vein, Clark (1975) distinguishes between bridging via necessary, probable, and inducible parts/roles. He argues that only in the first case does the antecedent trigger the bridging anaphor in the sense that we already think of the anaphor when we read/hear the antecedent. For instance, **walls** in Example (1) are necessary parts of the antecedent *the Polish center* according to common sense knowledge. However, windows and carpets are only probable or inducible parts of a building but still function as bridging anaphors in Example (1).

### 4.2 Method: Model

#### 4.2.1 Model I: Collective Classification

*Motivation.* Two mediated subcategories account for accessibility via syntactic links to another *old* or *mediated* mention. *Mediated/syntactic* is used when at least one child of a mention is *mediated* or *old*, with child relations restricted to:

- Possessive pronouns or possessive NPs  
(e.g.,  $[[\text{his}]_{\text{old}} \text{father}]_{\text{mediated/syntactic}}$ )
- Of-genitives (e.g.,  $[\text{The alcoholism of } [\text{his}]_{\text{old}} \text{father}]_{\text{mediated/syntactic}}$ )
- Proper name premodifiers  
(e.g.,  $[\text{The } [\text{Federal Reserve}]_{\text{mediated/WK}} \text{boss}]_{\text{mediated/syntactic}}$ )

<sup>14</sup> We thank an anonymous reviewer for bringing up this example.

- Other prepositional phrases  
(e.g., [professors at [Cambridge]<sub>mediated/WK</sub>]<sub>mediated/syntactic</sub>)

The subcategory *mediated/aggregate* is for coordinations in which at least one of the children is *old* or *mediated*, e.g., *Not only George Bush but also Barack Obama* is *mediated/aggregate* as *Barack Obama* is *mediated/WK*.

In these two cases, a mention's IS depends directly on the IS of its children. This is therefore a case of so-called **autocorrelation**, a characteristic of relational data in which the value of one variable for one instance is highly correlated with the value of the same variable on another instance. By exploiting relational autocorrelation, collective classification (Jensen, Neville, and Gallagher 2004; Macskassy and Provost 2007) can significantly outperform independent supervised classification (Taskar, Segal, and Koller 2001; Neville and Jensen 2003; Domingos and Lowd 2009) and has been applied, for example, in part-of-speech tagging (Lafferty, McCallum, and Pereira 2001), Web page categorization (Taskar, Abbeel, and Koller 2002), opinion mining (Somasundaran et al. 2009; Burfoot, Bird, and Baldwin 2011), and entity linking (Fahrni and Strube 2012).

*Detailed model.*  $M$  denotes the set of  $n$  mentions in a document  $D$ , and  $S$  the set of eight IS classes. Let  $s_m$  be the IS class associated with a mention  $m \in M$ ,  $S_M$  be the IS class assignments for all mentions in  $M$ ,  $S_M^n$  be the set of all possible IS class assignments for  $M$ . The collective IS classification task can be represented as a log-linear model:

$$P(S_M|M;w) = \frac{\exp(w \cdot \Phi(M, S_M))}{\sum_{S_M' \in S_M^n} \exp(w \cdot \Phi(M, S_M'))} \quad (17)$$

where  $w$  is the model's weight vector, and  $\Phi(M, S_M)$  is the feature vector that takes all IS class assignments for all mentions in  $M$  into account. We define  $\Phi(M, S_M)$  as:

$$\Phi(M, S_M) = \sum_{l \in F_l} \sum_{m \in M} \Phi_l(m, s_m) + \sum_{g \in F_g} \sum_{m_i, m_j \in M} \Phi_g(s_{m_i}, s_{m_j}) \quad (18)$$

where  $\Phi_l(m, s_m)$  is a local feature function that looks at the mention  $m$  and the target IS class  $s_m$ ,  $F_l$  is the set of local features,  $\Phi_g(s_{m_i}, s_{m_j})$  is a global feature function that looks at the target IS class assignments for  $m_i$  and  $m_j$  at once, and  $F_g$  is the global feature set.

This log-linear model can be represented using Markov logic networks (MLNs) (Domingos and Lowd 2009). An MLN is a statistical relational learning framework that combines *first order logic* and *Markov networks*. It provides us with a simple yet flexible language to construct joint models for bridging resolution. Moreover, our task-specific models can benefit from the advances in inference and learning algorithms for MLNs.

A Markov logic network is defined as a set of pairs  $(f_i, w_i)$ , where  $f_i$  is a formula in first-order logic and  $w_i$  is a real number (Domingos and Lowd 2009). In first-order logic, formulas are constructed using four types of symbols: *constants*, *variables*, *functions*, and *predicates*. Constant symbols represent objects that we are interested in (mentions in our problem, such as *his father* or *his*), variable symbols range over objects in the domain, function symbols map objects to objects, and predicate symbols represent relations among objects or attributes of objects (e.g., *hasIS* in Table 5).

**Table 5**

Hidden predicates and formulas used for bridging anaphora recognition.  $m$  represents a mention,  $M$  the set of mentions in the whole document,  $s$  an IS class,  $S$  the set of eight IS classes, and  $w$  the weight learned from the data for the specific formula.

**Hidden predicates**

$p1$   $hasIS(m, s)$

**Formulas**

Hard constraints

$f1$   $\forall m \in M : |s \in S : hasIS(m, s)| = 1$

Joint inference formula template

$f_g$   $(w) \quad \forall m_i, m_j \in M \forall s_{m_i}, s_{m_j} \in S : \text{jointInferenceFormula.Constraint}(m_i, m_j) \rightarrow hasIS(m_i, s_{m_i}) \wedge hasIS(m_j, s_{m_j})$

Non-joint inference formula template

$f_l$   $(w) \quad \forall m \in M \forall s \in S : \text{non-jointInferenceFormula.Constraint}(m, s) \rightarrow hasIS(m, s)$

In a ground Markov network, the probability distribution over the possible world  $S_M$  is given by

$$P(S_M) = \frac{1}{Z} \exp \left( \sum_i w_i n_i(S_M) \right) \quad (19)$$

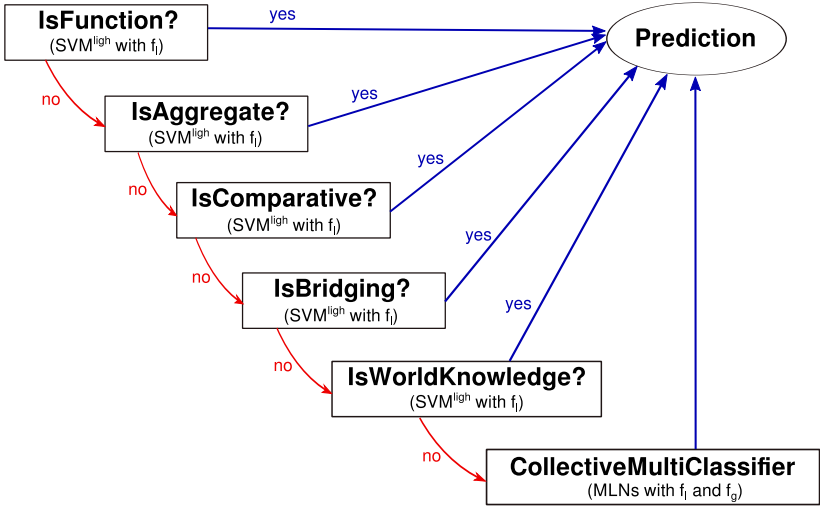
where  $n_i(S_M)$  is the number of true groundings of a local or a global feature function  $F_i$  in  $S_M$ ,  $w_i$  is the weight for  $F_i$ , and  $Z$  is the partition function. Table 5 shows the formula templates to model IS recognition in MLNs.  $p1$  is the hidden predicate that we want to predict (i.e., the information status  $s$  of a mention  $m$ ).  $f1$  models that each mention can only belong to one IS class, and  $f_g$  and  $f_l$  are templates of joint inference formulas and non-joint inference formulas, respectively.<sup>15</sup> Details of specific formulas (features) instantiating  $f_g$  and  $f_l$  are described in Sections 4.3.1 and 4.3.2.

#### 4.2.2 Model II: Cascading Collective Classification

*Motivation for the model.* As shown in Section 3, bridging anaphors have different determiners and few easily identifiable surface features. In addition, they are relatively rare, making up only 6% of noun phrases in our data. Such multi-class imbalance problems are an open research topic (Abe, Zadrozny, and Langford 2004; Zhou and Liu 2010; Wang and Yao 2012). Classification accuracy may be artificially high in case of extremely imbalanced data: Majority classes are favored, and minority classes are not recognized. Such a bias becomes stronger within the multi-class setting. To address this problem while still keeping the strength of collective inference within a multi-class setting, we

<sup>15</sup>  $f_g$  and  $f_l$  correspond to  $F_g$  and  $F_l$ , respectively, in Equation (18).





**Figure 3**  
The cascading collective classification system.

integrate our collective classification model (Section 4.2.1) into a cascading collective classification system inspired by Omuya, Prabhakaran, and Rambow (2013).

*Detailed Model.* Unlike in the multi-class setting, learning from imbalanced data in the binary setting has been well studied (He and Garcia 2009). Therefore, our cascading collective classification system, shown in Figure 3, combines binary classifiers for minority categories and a collective classifier for all categories in a pipeline. Specifically, for the five classes mediated/function, mediated/aggregate, mediated/comparative, mediated/bridging, and mediated/WK that each constitutes less than the expected one-eighth of the instances, we develop five binary classifiers with SVM<sup>light</sup> (Joachims 1999). These classifiers use only non-joint inference formulae, but have the advantage that we can tune the SVM parameter against data imbalance on the training set. We arrange them from the rarest to the most frequent category. Whenever a minority classifier predicts true, this class is assigned. When all minority classifiers say false, we back off to multi-class collective inference (Section 4.2.1). Omuya, Prabhakaran, and Rambow (2013, page 805) motivate a rarest to most frequent ordering on the task of dialogue act tagging by “the observation that the less frequent classes are also hard to predict correctly” and we follow their procedure. Such a framework substantially improves bridging anaphora recognition without jeopardizing performance on other IS classes (Section 4.4.3).

**4.3 Method: Features**

Section 4.3.1 details the relational features that instantiate the joint inference formula template  $f_g$  in Table 5. Section 4.3.2 details non-relational features that instantiate the non-joint inference formula template  $f_i$  in Table 5. Apart from the ISNotes corpus, for some non-relational features we use additional resources with manual annotation, namely, NomBank (Meyers et al. 2004), WordNet, the General Inquirer (Stone et al. 1966), and the ACE2 annotations for genericity (Mitchell et al. 2002).

### 4.3.1 Relational Features

*Syntactic hasChild relations.* We 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)  $m_2$  is a proper name premodifier of  $m_1$ . For instance, the mention [professors at Cambridge] is linked to the mention [Cambridge] via a *hasChild* relation.

*Syntactic hasChildCoordination relations.* We link a mention  $m_1$  to a mention  $m_2$  via a *hasChildCoordination* relation if  $m_1$  is a coordination and  $m_2$  is one of its children. For example, the mention [Not only George Bush but also Barack Obama] is linked to the mention [Barack Obama] via a *hasChildCoordination* relation.

*Syntactic ConjoinedTo relations.* Conjoined mentions may have the same IS class. We link a mention  $m_1$  to a mention  $m_2$  via a *ConjoinedTo* relation if both  $m_1$  and  $m_2$  are the children of a coordination. For example, [George Bush] is linked to [Barack Obama] via a *ConjoinedTo* relation as both are the children of the coordination [Not only George Bush but also Barack Obama].

**4.3.2 Non-relational Features.** Table 6 shows all non-relational features for IS recognition.

*Features p1–p8 from previous work.* We adapt features p1–p8 from Nissim (2006) and Rahman and Ng (2011). A mention with complete string match to a previous one is likely to be old (p1, p2). The head match feature p3 (from Nissim’s PartialPreMention feature as well as coreference resolution [Vieira and Poesio 2000; Soon, Ng, and Lim 2001]) identifies old and mediated categories such as comparative anaphora. p4 *NPlength* is motivated by Arnold et al. (2000, page 34): “items that are new to the discourse tend to be complex and items that are given tend to be simple.” There is a tendency for indefinite NPs to be new (Hawkins 1978) (p5). Subjects are likely to be old (p6) (Prince 1992). Pronouns tend to be old (p7). Rahman and Ng (2011) explore lexical features (p8), for example, mentions which include the lexical unit *his* are likely not to be new.

*New features for identifying several IS classes (non-bridging).* The new features (g1–g5) capture the classes old as well as mediated/WK, mediated/comparative, and mediated/function. g1 *HeadMatchTime* and g2 *ContentWordPreMention* are string match variations, giving a categorical version of p3 *HeadMatch* and a partial mention match going beyond the mention’s head, respectively.

Proper names not previously mentioned in the text but appearing in many other documents are likely to be hearer-old (IS class mediated/WK). To approximate this, g3 *IsFrequentProperName* measures if the mention is a proper name, occurring in at least 100 documents in the Tipster corpus (Harman and Liberman 1993).

Mediated/comparative mentions are often indicated by surface clues such as premodifiers (e.g., *other*, *another*). In g4 *PreModByCompMarker*, we check for such markers<sup>16</sup> as well as the presence of adjectives or adverbs in the comparative form.

<sup>16</sup> The full list is: {*other*, *another*, *such*, *different*, *similar*, *additional*, *comparable*, *same*, *further*, *extra*}.

**Table 6**  
Non-relational features for IS classification. “b” indicates binary features, “n” nominal features, “l” lexical features, “int” integer. A nominal feature draws the feature value from a restricted set. A lexical feature indicates the presence or absence of a lexical unit in a mention. The value “NA” stands for “not applicable” and is used for pronouns.

Feature	Value
<b>Features from previous work (Nissim 2006; Rahman and Ng 2011)</b>	
<i>p1</i> FullPrevMention (n)	{ <i>yes, no, NA</i> } <sup>1</sup>
<i>p2</i> FullMentionTime (n)	{ <i>first, second, more, NA</i> }
<i>p3</i> HeadMatch (n)	{ <i>yes, no, NA</i> }
<i>p4</i> NPlength (int)	numeric, e.g., 5
<i>p5</i> Determiner (n)	{ <i>def, indef, dem, poss, bare, NA</i> }
<i>p6</i> GrammaticalRole (n)	{ <i>subject, subypass, object, predicate, pp, other</i> }
<i>p7</i> NPType (n)	{ <i>common noun, proper noun, pronoun, other</i> }
<i>p8</i> Unigrams (l)	e.g., <i>his, the, China</i>
<b>New features for identifying several IS classes (non-bridging)</b>	
<i>g1</i> HeadMatchTime (n)	{ <i>first, second, more, NA</i> }
<i>g2</i> ContentWordPreMention (b)	{ <i>yes, no, NA</i> }
<i>g3</i> IsFrequentProperName (b)	{ <i>yes, no</i> }
<i>g4</i> PreModByCompMarker (b)	{ <i>yes, no</i> }
<i>g5</i> DependOnChangeVerb (b)	{ <i>yes, no</i> }
<b>New features for recognizing bridging anaphora</b>	
<i>Discourse structure</i>	
<i>f1</i> IsCoherenceGap (b)	{ <i>yes, no</i> }
<i>f2</i> IsSentFirstMention (b)	{ <i>yes, no</i> }
<i>f3</i> IsDocFirstMention (b)	{ <i>yes, no</i> }
<i>Lexico-semantics</i>	
<i>f4</i> IsArgumentTakingNP (b)	{ <i>yes, no</i> }
<i>f5</i> IsWordNetRelationalNoun (b)	{ <i>yes, no</i> }
<i>f6</i> IsInquirerRoleNoun (b)	{ <i>yes, no</i> }
<i>f7</i> SemanticClass (n)	a list of 16 classes, e.g., <i>location, organization</i>
<i>f8</i> IsBuildingPart (b)	{ <i>yes, no</i> }
<i>f9</i> IsSetElement (b)	{ <i>yes, no</i> }
<i>f10</i> ModSpatialTemporal (b)	{ <i>yes, no</i> }
<i>f11</i> IsYear (b)	{ <i>yes, no</i> }
<i>f12</i> PreModifiedByCountry (b)	{ <i>yes, no</i> }
<i>Identifying generic NPs</i>	
<i>f13</i> AppearInIfClause (b)	{ <i>yes, no</i> }
<i>f14</i> NPNumber (n)	{ <i>singular, plural, unknown</i> }
<i>f15</i> VerbPosTag (l)	e.g., <i>VBG, MD, VB</i>
<i>f16</i> IsFrequentGenericNP (b)	{ <i>yes, no</i> }
<i>f17</i> GeneralWorldKnowledge(l)	e.g., <i>the sun, the wind</i>
<i>f18</i> PreModByGenericQuantifier (b)	{ <i>yes, no</i> }
<i>Mention syntactic structure</i>	
<i>f19</i> HasChildMention (b)	{ <i>yes, no</i> }

<sup>1</sup> We changed the value of “*f1* FullPrevMention” from “numeric” to {*yes, no, NA*}.

*g5 DependOnChangeVerb* determines whether a number mention is the object of an increase/decrease verb and therefore is likely to be the IS class *mediated/function*.<sup>17</sup>

*New features for recognizing bridging anaphors.* Bridging anaphors are rarely marked by surface features but are often licensed because of discourse structure and/or lexical or world knowledge. Motivated by these observations, we develop discourse structure and lexico-semantic features indicating bridging anaphora. We also design features to separate genericity from bridging anaphora.

Discourse structure features (Table 6, *f1–f3*). Bridging is sometimes the only means to establish entity coherence to previous sentences/clauses (Grosz, Joshi, and Weinstein 1995; Poesio et al. 2004b). This is especially true for *topic* NPs (Halliday and Hasan 1976). We therefore define *coherence gap sentences* as sentences that have none of the following three coherence elements: (1) entity coreference to previous sentences, as approximated via string match or the presence of pronouns; (2) comparative anaphora approximated by mentions modified via 10 comparative markers, or by the presence of adjectives or adverbs in the comparative (see also *g4 PreModByCompMarker*); or (3) proper names.<sup>18</sup> Bridging Examples (1), (9), (10), (11), (12), (14), and (16) occur in coherence gap sentences under our definition. We approximate the topic of a sentence via the first mention (*f2 IsSentFirstMention*). *f3 IsDocFirstMention* models that bridging anaphors do not appear at the beginning of a text.

Lexico-semantic features (Table 6, *f4–f12*). Drawing on theories of noun types (Löbner 1985) and bridging sub-classes (Clark 1975; Poesio and Vieira 1998; Lassalle and Denis 2011), we capture lexical properties of head nouns of bridging.

Löbner (1985) distinguishes between relational nouns that take on at least one core semantic role (such as *friend*) and sortal nouns (such as *table* or *flower*). He points out that relational nouns are more frequently used for bridging than sortal nouns (see Examples (8), (9), (13), and (14)). *f4 IsArgumentTakingNP* and *f5 IsWordNetRelationalNoun* capture relational nouns. *f4* decides whether the argument taking ratio of a mention's head is bigger than some threshold *k*. We calculate the argument taking ratio  $\alpha$  for a mention using NomBank (Meyers et al. 2004). For each mention,  $\alpha$  is calculated via its head frequency in the NomBank annotation divided by the head's total frequency in the WSJ corpus on which the NomBank annotation is based. The value of  $\alpha$  reflects how likely an NP is to take arguments. For instance, the value of  $\alpha$  is 0.90 for *husband* but 0.31 for *children*. We also extract around 4,000 relational nouns from WordNet, then determine whether the mention head appears in the list or not (*f5 IsWordNetRelationalNoun*). The core semantic role for a relational noun can of course also be filled NP-internally instead of anaphorically. We use the features *f12 PreModifiedByCountry* (such as *the Egyptian president*) and *f19 HasChildMention* (for complex NPs that are likely to fill needed roles NP-internally) to address this.

Role terms (e.g., *chairman*) and kinship terms (e.g., *husband*) are also relational nouns. *f6 IsInquirerRoleNoun* determines whether the mention head appears under the *role* category in the General Inquirer lexicon (Stone et al. 1966). The feature *f7 Semantic-Class* puts each mention into one of 16 coarse-grained semantic classes: {rolePerson, relativePerson, person\*, organization, geopolitical entity (GPE), location, nationality or

17 We extract increase/decrease verbs from the General Inquirer lexicon (Stone et al. 1966). The list contains the verbs {*increase, raise, rise, climb, swell, ascend, jump, leap, scale, stretch, become, double, extend, grow, improve, strengthen, fall, drop, cut, slow, ease, reduce, descend, lower, slip*}.

18 Note that we use the notion of a coherence gap as missing entity coherence to all previous sentences, not just the adjacent one as discussed in Grosz, Joshi, and Weinstein (1995).

religious or political group (NORP), event, product, date, time, percent, money, ordinal, cardinal, other}, using the OntoNotes annotation for named entities and WordNet for common nouns. The category *rolePerson* matches person mentions whose head noun specifies a professional role such as *mayor*, *director*, or *president*, using a list of 100 such nouns from WordNet. The category *relativePerson* matches person mentions whose head noun specifies a family or friend role such as *husband*, *daughter*, or *friend*, using a list of 100 such nouns from WordNet. The category *person\** is assigned to all other person mentions.

Because part-of relations are typical bridging relations (see Example (1) and Clark [1975]), *f8 IsBuildingPart* determines whether the mention head might be a part of a building, using a list of 45 nouns from the General Inquirer under the *BldgPt* category.

*f9 IsSetElement* is used to identify set-membership bridging cases (see Example (12)), by checking whether the mention head is a number or indefinite pronoun (*one*, *some*, *none*, *many*, *most*) or modified by *each*, *one*. However, not all numbers are bridging cases, and we use *f11 IsYear* to exclude some such cases.

Some bridging anaphors are indicated by spatial or temporal modifiers (see Example (11) and also Lassalle and Denis [2011]). We use *f10 ModSpatialTemporal* to detect these cases by compiling 22 such modifiers from the General Inquirer (Stone et al. 1966).<sup>19</sup>

Features to detect generic NPs (Table 6, *f13–f18*). Generic NPs (Example (15)) are easily confused with bridging. Inspired by Reiter and Frank (2010), we develop features (*f13–f18*) to exclude generics.

First, hypothetical entities are likely to refer to generic entities (Mitchell et al. 2002). We approximate this by determining whether the NP appears in an if-clause (*f13 AppearInIfClause*). Also the NP's number (e.g., *singular* or *plural*) and the clause tense/mood may play a role to decide genericity (Reiter and Frank 2010). The former is detected on the basis of the POS tag of the mention's head word (*f14 NPNumber*). The latter is often reflected by the verb form of the clause where the mention is present, such as *VBG* or *MD VB VBG*. So we use the POS tags of the clause verbs as lexical features (*f15 VerbPosTag*).

The ACE-2 corpus (Mitchell et al. 2002) (distinct from our corpus) contains annotations for genericity. We collect all NPs from ACE-2 that are always used generically (*f16 IsFrequentGenericNP*). We also try to learn NPs that are uniquely identifiable without further description or anaphoric links such as *the sun* or *the pope*, by extracting common nouns that are annotated as *mediated/WK* from the training set and use these as lexical features (*f17 GeneralWorldKnowledge*).

Motivated by the ACE-2 annotation guidelines, *f18 PreModByGenericQuantifier* identifies six quantifiers that may indicate genericity (*all*, *no*, *neither*, *every*, *any*, *most*).

## 4.4 Results and Discussion

**4.4.1 Experimental Set-up.** Because of the still limited size of our annotated corpus, especially for the rarer IS categories, we conduct experiments via document-wise 10-fold cross-validation. We use the OntoNotes named entity and syntactic annotation for feature extraction. The value of the parameter *k* in the feature *f4 IsArgumentTakingNP* (Table 6) is estimated for each fold separately: We first choose ten

<sup>19</sup> The whole list is: {*final*, *first*, *last*, *next*, *prior*, *succeeding*, *second*, *nearby*, *previous*, *close*, *above*, *adjacent*, *behind*, *below*, *bottom*, *early*, *formal*, *future*, *before*, *after*, *earlier*, *later*}.

documents randomly from the training set for each fold as the development set to estimate  $k$ 's value via a grid search over  $k \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ , then the whole training set is trained again using the optimized parameter. We use recall, precision, and F-score to measure the performance per category. Accuracy measures overall performance on all IS categories. Statistical significance is measured using McNemar's  $\chi^2$  test (McNemar 1947).

*4.4.2 Evaluation of New Non-relational Features.* We reimplemented the two local IS classifiers in Nissim (2006) and Rahman and Ng (2011) as baselines (henceforth *Nissim* and *RahmanNg*), using their feature and algorithm choices. We then add our new features from Table 6 in Section 4.3.2 to the two baselines, yielding the following six systems.

*Nissim.* Algorithm *Nissim* is a J48 decision tree with standard settings in WEKA (Witten and Frank 2005), using features  $p1$ – $p7$  from Table 6.  $p8$  is not used in *Nissim*, because lexical features are not handled well by J48 (Nissim 2006).

*Nissim plus  $g1$ – $g5$ .* Features  $g1$ – $g5$  from Table 6 are added to *Nissim*. These new features are designed for the categories *old*, *mediated/WK*, *mediated/comparative*, and *mediated/function*.

*Nissim plus  $g1$ – $g5$  plus  $f1$ – $f19$ .* Features  $f1$ – $f19$  from Table 6 designed for *mediated/bridging* are added. As for algorithm *Nissim*, we again exclude lexical features ( $f15$  *VerbPosTag* and  $f17$  *GeneralWorldKnowledge*).

*RahmanNg.* Rahman and Ng (2011) use a binary SVM with a composite kernel (Joachims 1999; Moschitti 2006) on the Switchboard corpus. They use the one-versus-all strategy for multi-class classification and the features  $p1$ – $p8$  from Table 6. In addition, they use a tree kernel feature where the context of a mention is represented by its parent and its sibling nodes (without lexical leaves). Although this feature captures the syntactic context of a mention, it does not capture the internal structure of the mention itself nor the interaction between the IS of a mention and its children or parents.

*RahmanNg plus  $g1$ – $g5$ .* Features  $g1$ – $g5$  from Table 6 are added.

*RahmanNg plus  $g1$ – $g5$  plus  $f1$ – $f19$ .* Features  $f1$ – $f19$  (Table 6) are added.

Results for adding the new features to *Nissim* are shown in Table 7 (top) and to *RahmanNg* in Table 7 (bottom). The final algorithm improves significantly over all previous models in overall accuracy, showing the effectiveness of our new features. Comparative anaphors are recognized reliably via a small set of comparative markers. Including features  $g3$  *IsFrequentProperName* and  $g5$  *DependOnChangeVerb* improves results for *mediated/WK* and *mediated/function*, respectively.

Features  $f1$ – $f19$  from Table 6 were specifically designed for bridging: They help *Nissim plus  $g1$ – $g5$  plus  $f1$ – $f19$*  improve the results for bridging substantially over *Nissim plus  $g1$ – $g5$* . They also help to delimit several other IS classes better, such as *mediated/syntactic* for *Nissim plus  $g1$ – $g5$  plus  $f1$ – $f19$*  and *RahmanNg plus  $g1$ – $g5$  plus  $f1$ – $f19$* .

The new features  $f1$ – $f19$  only have limited effect on bridging recognition in *RahmanNg plus  $g1$ – $g5$  plus  $f1$ – $f19$*  compared with *RahmanNg*. Unigrams in *RahmanNg*

**Table 7**  
Experimental results: compared to the baseline *Nissim* (top) and *Rahman* (bottom). **Bolded** scores indicate significant improvements relative to all other models ( $p < 0.01$ ).

	<i>Nissim</i>			<i>Nissim plus g1-g5</i>			<i>Nissim plus g1-g5 plus f1-f19</i>		
	R	P	F	R	P	F	R	P	F
old	85.1	82.7	83.9	85.6	82.5	84.0	85.6	85.4	85.5
med/worldKnowledge	62.3	64.4	63.3	64.2	72.0	67.8	63.3	76.3	69.2
med/syntactic	41.6	59.7	49.0	44.8	61.8	52.0	59.2	63.9	<b>61.4</b>
med/aggregate	28.4	36.8	32.1	31.8	44.7	37.1	34.6	44.5	38.9
med/function	0.0	NA	NA	38.5	89.3	53.8	58.5	76.0	<b>69.2</b>
med/comparative	0.4	7.7	0.7	84.6	82.0	83.3	83.0	78.1	80.5
med/bridging	<b>4.4</b>	<b>23.0</b>	<b>7.4</b>	<b>5.3</b>	<b>24.5</b>	<b>8.9</b>	<b>20.7</b>	<b>41.5</b>	<b>27.6</b>
new	82.7	62.3	71.1	82.0	65.4	72.8	79.7	68.7	73.8
Accuracy	67.6			70.4			<b>72.6</b>		
	<i>RahmanNg</i>			<i>RahmanNg plus g1-g5</i>			<i>RahmanNg plus g1-g5 plus f1-f19</i>		
	R	P	F	R	P	F	R	P	F
old	85.3	87.1	86.2	85.7	86.9	86.3	86.8	86.7	86.8
med/worldKnowledge	66.6	69.6	68.0	67.1	73.5	70.1	64.9	81.2	<b>72.2</b>
med/syntactic	57.3	72.2	63.9	55.8	72.8	63.2	66.3	71.7	<b>68.9</b>
med/aggregate	26.5	75.7	39.3	25.1	73.6	37.5	29.4	78.5	<b>42.8</b>
med/function	24.6	51.6	33.3	56.9	84.1	67.9	44.6	85.3	58.6
med/comparative	26.5	85.9	40.5	79.4	81.7	80.6	79.1	81.0	80.0
med/bridging	<b>11.6</b>	<b>45.6</b>	<b>18.5</b>	<b>8.9</b>	<b>44.7</b>	<b>14.8</b>	<b>12.4</b>	<b>61.2</b>	<b>20.6</b>
new	87.8	66.7	75.8	87.6	67.6	76.3	87.4	70.0	<b>77.8</b>
Accuracy	73.3			74.4			<b>76.2</b>		

may cover the lexical knowledge for bridging anaphora recognition that we model explicitly via features. Also although the overall IS classification performance of *RahmanNg plus g1-g5 plus f1-f19* is significantly better than *Nissim plus g1-g5 plus f1-f19*, the former is worse than the latter with regard to bridging anaphora recognition. The one-versus-all strategy for a multi-class setting in Rahman and Ng (2011) is not suitable for identifying a minority class which lacks strong indicators such as bridging.

*4.4.3 Evaluation of Collective and Cascaded Collective Classification.* We now compare the best local classifier, *RahmanNg plus g1-g5 plus f1-f19*, to collective and cascaded collective classifiers (*Collective* and *CascadedCollective*). The MLN classifier *Collective* (Section 4.2.1) uses the non-relational features from Table 6 and adds the relational features from Section 4.3.1. We use *thebeast*<sup>20</sup> to learn weights and to perform

<sup>20</sup> <http://code.google.com/p/thebeast>.

**Table 8**  
Experimental results: Comparing the best local to collective and cascaded collective classifiers. **Bolded** scores indicate significant improvements compared to previous model ( $p < 0.01$ ).

	<i>RahmanNg plus g1–g5 plus f1–f19</i>			<i>Collective</i>			<i>CascadedCollective</i>		
	R	P	F	R	P	F	R	P	F
old	86.8	86.7	86.8	85.7	84.7	85.2	83.1	85.7	84.4
med/worldKnowledge	64.9	81.2	72.2	64.2	80.5	71.4	65.6	79.5	71.9
med/syntactic	66.3	71.7	68.9	82.7	80.1	<b>81.4</b>	82.2	80.5	81.3
med/aggregate	29.4	78.5	42.8	71.6	78.2	<b>74.8</b>	71.1	77.7	74.3
med/function	44.6	85.3	58.6	56.9	90.2	69.8	61.5	83.3	70.8
med/comparative	79.1	81.0	80.0	81.4	84.8	83.1	83.4	82.7	83.1
med/bridging	12.4	61.2	20.6	25.9	49.9	<b>34.1</b>	48.7	43.8	<b>46.1</b>
new	87.4	70.0	77.8	84.5	75.7	<b>79.9</b>	81.3	77.7	79.5
Accuracy	76.2			<b>78.9</b>			78.4		

inference.<sup>21</sup> *thebeast* uses cutting plane inference (Riedel 2008) to improve the accuracy and efficiency of MAP inference for MLNs.

The relational features in *Collective* lead to significant improvements in accuracy over the local model (Table 8), in particular for mediated/syntactic and mediated/aggregate as well as their distinctions from new. Such improvement is in accordance with the linguistic relations among IS categories we analyzed in Section 4.2.1.<sup>22</sup> *Collective* also improves the F-score for bridging by 13.5% compared with the local model. This is mainly through improved recall, where the local model in Table 8 is very conservative with a recall score of only 12.4%. *Collective* doubles recall but at a certain loss to precision.

However, the results for the bridging category, including recall, are still low. In a multi-class setting, prediction is biased toward the classes with the highest priors. *CascadedCollective* classification (Section 4.2.2) addresses this problem by combining a sequence of minority binary classifiers (based on SVMs, using only non-relational features) with a final collective classifier (based on MLNs, using non-relational and relational features). *CascadedCollective* improves bridging F-score and recall substantially without jeopardizing performance on other IS classes (Table 8, right). One question is whether the cascading algorithm is sufficient for improved bridging recognition with our additional non-relational bridging features f1–f19 being superfluous. We ran *CascadedCollective* without these features. Results worsened substantially to 74.4% overall accuracy and 29.2 bridging F-measure. Our novel features (addressing linguistic properties of bridging) and the cascaded algorithm (addressing data sparseness) are complementary.

21 In 10-fold cross-validation, we have 45 training documents in each fold. In fold0, the ground Markov network of *Collective* for the first training instance contains 5,831 variables and it takes around 35 minutes on an 8 CPU core machine to train the model.  
22 The improvement is not due simply to a switch from SVMs to MLNs. If we run the MLN without the novel relational features, we obtain performance comparable but slightly lower than SVMs.



**Table 9**  
Confusion matrix of *CascadedCollective* for bridging anaphora recognition. “C” indicates classifier tags, “G” gold tags. “brid” stands for *mediated/bridging*, “syn” *mediated/syntactic*, “comp” *mediated/comparative*, “aggr” *mediated/aggregate*, “func” *mediated/function*, “know” *mediated/worldKnowledge*.

C → G ↓	old	new	brid	syn	comp	aggr	func	know
old	-	-	175	-	-	-	-	-
new	-	-	193	-	-	-	-	-
brid	66	251	323	10	2	1	0	10
synt	-	-	10	-	-	-	-	-
comp	-	-	2	-	-	-	-	-
aggr	-	-	0	-	-	-	-	-
func	-	-	0	-	-	-	-	-
know	-	-	35	-	-	-	-	-

**4.4.4 Error Analysis.** Our performance on bridging recognition, although outperforming reimplementations of previous work, is still under 50% in all measures. We conducted an error analysis using our best model *CascadedCollective*. We examine the confusion matrix (Table 9) of the model, concentrating only on the numbers related to bridging.

The highest proportion of recall errors is due to the fact that 251 bridging anaphors are misclassified as new. This can be explained as the syntactic form of many new instances and bridging anaphors are the same, new items are more frequent, and our lexico-semantic features in particular only pick up on certain types of bridging.

Most precision errors are new and old instances being misclassified as mediated/bridging. Many old instances misclassified as bridging are definite NPs without further modification and common noun heads without a string match to a previous mention. An example would be an NP such as *the president*, which can easily be coreferent to a previous president named by proper name (*Barack Obama*) or a bridging to a country or company. This coincides with the fact that in coreference resolution, common noun anaphors without head match are also hardest to detect (Martschat and Strube 2014). Future work attempting joint bridging and coreference resolution might help here. New items misclassified as bridging are also NPs with common noun heads and no modification (outside determiners) such as *control* or *the back*, often generics (see Examples (14) and (15)). In the latter cases how the phrase is embedded in the discourse plays an important role and is only partially modeled by our approach. Currently, the lexical semantic knowledge we explored only indicates that some NPs are more likely to be used as bridging anaphora than others.

**5. Bridging Antecedent Selection**

Bridging antecedent selection chooses an antecedent among all possible candidates for a given bridging anaphor. We make contributions in three areas for antecedent selection: (i) using joint modeling to tackle what we call sibling anaphora, (ii) developing a range of semantic and salience features for the problem, and (iii) proposing the novel concept of an anaphor’s discourse scope to delimit the list of possible candidate antecedents.

From now on we assume that the antecedent is an NP mention—because, among 663 bridging anaphors, only 39 have verbs/clauses as antecedents (see Section 3).

Downloaded from [http://direct.mit.edu/colli/article-pdf/44/2/237/1808960/colli\\_a\\_00315.pdf](http://direct.mit.edu/colli/article-pdf/44/2/237/1808960/colli_a_00315.pdf) by guest on 31 March 2021

We do not resolve the latter and count our decisions in these cases as incorrect. The antecedent can be coreferent with prior mentions of the same entity. In Example (1), repeated as Example (20), **The windows** is the bridging anaphor, *the Polish center* is the antecedent (mention), and the antecedent is coreferent to *the center* mentioned previously. We call such a coreference chain of antecedents the *antecedent entity*.

If we see bridging as independent of coreference resolution, we can select antecedents among all prior mentions—if not, then among all entities mentioned before the anaphor. The first case can be seen as a special case of the second one where all antecedent entities and candidates are chains of length 1 and we know nothing about their coreference properties. For our general model formulation, we assume the antecedent entity setting as the general case which subsumes the other setting. However, candidate generation, feature computation, and evaluation will vary for the two settings and therefore we explore both scenarios in the experiments.

- (20) A cake topped with a replica of *the center* will be auctioned at an AIDS benefit at Sotheby's in December. If Mr. McDonough's plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. [2 sent.] **The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

### 5.1 Method: A Joint Model

*Motivation.* Many of our bridging anaphors are siblings—that is, they share the same antecedent (Section 3). *Sibling anaphors clustering* tries to identify such siblings. We then use joint inference to model *sibling anaphors clustering* and *bridging antecedent selection* together.

*Detailed model.*  $A$  denotes the set of  $n$  bridging anaphors in document  $D$ .  $E$  denotes the set of antecedent candidates in the whole document. Let  $c_{a_i/a_j}$  be a sibling anaphors clustering assignment for bridging anaphors  $a_i, a_j \in A$ ,  $C_A$  be a sibling anaphors clustering result for all bridging anaphors in  $A$ , and  $C_A^n$  be the set of all possible sibling anaphors clustering results for  $A$ . Let  $e_a$  be an antecedent assignment for a bridging anaphor  $a \in A$ ,  $E_A$  be an antecedent assignment result for all bridging anaphors in  $A$ , and  $E_A^n$  be the set of all possible antecedent assignment results for  $A$ . Joint inference for *sibling anaphors clustering* and *bridging antecedent selection* can be represented as a log-linear model:

$$P(C_A, E_A | A; w) = \frac{\exp(w \cdot \Phi(A, C_A, E_A))}{\sum_{E_A' \in E_A^n, C_A' \in C_A^n} \exp(w \cdot \Phi(A, C_A', E_A'))} \quad (21)$$

where  $w$  is the model's weight vector,  $\Phi(A, C_A, E_A)$  is a "global" feature vector that takes the entire clustering and antecedent assignments for all bridging anaphors in  $A$  into account. We define  $\Phi(A, C_A, E_A)$  as:

$$\begin{aligned} \Phi(A, C_A, E_A) = & \sum_{l \in F_c} \sum_{a_i, a_j \in A} \Phi_l(a_i, a_j, c_{a_i/a_j}) + \sum_{k \in F_r} \sum_{a \in A} \Phi_k(a, e_a) \\ & + \sum_{g \in F_g} \sum_{a_i, a_j \in A} \Phi_g(c_{a_i/a_j}, e_{a_i}, e_{a_j}) \end{aligned} \quad (22)$$

where  $\Phi_l(a_i, a_j, c_{a_i/a_j})$  and  $\Phi_k(a, e_a)$  are local feature functions for *sibling anaphors clustering* and *bridging antecedent selection*, respectively. The former looks at two bridging anaphors  $a_i$  and  $a_j$ , the latter at the bridging anaphor  $a$  and the antecedent candidate  $e_a$ .  $F_c$  and  $F_r$  are the sets of local features for these two tasks, respectively. The global feature function  $\Phi_g(c_{a_i/a_j}, e_{a_i}, e_{a_j})$  looks at the antecedent assignments for  $a_i$  and  $a_j$  at the same time, and  $F_g$  is the set of global features.

Like our collective classification model (Section 4.2.1), this log-linear model can be represented using MLNs. In a ground Markov network for this task, the probability distribution over the possible world  $C_A, E_A$  is given by

$$P(C_A, E_A) = \frac{1}{Z} \exp \left( \sum_i w_i n_i(C_A, E_A) \right) \quad (23)$$

where  $n_i(C_A, E_A)$  is the number of true groundings of a local or a global feature function  $F_i$  in  $(C_A, E_A)$ ,  $w_i$  is the weight for  $F_i$ , and  $Z$  is the partition function.

Table 10 shows hard constraints and formula templates for this problem in MLNs.

$p1$  and  $p2$  are hidden predicates that we predict: choosing the antecedent for anaphor  $a1$  and deciding whether  $a_1$  and  $a_2$  are sibling anaphors.  $f1$  models that each

**Table 10**

Hidden predicates, hard constraints, and formula templates used for bridging antecedent selection.  $a_1, a_2, a_3$  represent bridging anaphors,  $A$  the set of bridging anaphors in the whole document,  $e$  the antecedent candidate,  $E_a$  the set of the antecedent candidates for  $a$  according to  $a$ 's discourse scope, and  $E$  the set of antecedent candidates in the whole document.

---

**Hidden predicates**

---

- $p1$      $isBridging(a_1, e)$   
 $p2$      $hasSameAntecedent(a_1, a_2)$
- 

**Hard Constraints**

---

- $f1$      $\forall a \in A : |e \in E : isBridging(a, e)| \leq 1$   
 $f2$      $\forall a \in A \forall e \in E : hasPairDistance(e, a, d) \wedge d < 0 \rightarrow \neg isBridging(a, e)$   
 $f3$      $\forall a_1, a_2 \in A : a_1 \neq a_2 \wedge hasSameAntecedent(a_1, a_2) \rightarrow hasSameAntecedent(a_2, a_1)$   
 $f4$      $\forall a_1, a_2, a_3 \in A : a_1 \neq a_2 \wedge a_1 \neq a_3 \wedge a_2 \neq a_3 \wedge hasSameAntecedent(a_1, a_2)$   
           $\wedge hasSameAntecedent(a_2, a_3) \rightarrow hasSameAntecedent(a_1, a_3)$   
 $f5$      $\forall a_1, a_2 \in A \forall i \in E : a_1 \neq a_2 \wedge hasSameAntecedent(a_1, a_2) \wedge isBridging(a_1, e)$   
           $\rightarrow isBridging(a_2, e)$   
 $f6$      $\forall a_1, a_2 \in A \forall e \in E : a_1 \neq a_2 \wedge isBridging(a_1, e) \wedge isBridging(a_2, e)$   
           $\rightarrow hasSameAntecedent(a_1, a_2)$
- 

Formula template for sibling anaphors clustering

- $f_c$      $\forall a_1, a_2 \in A : siblingAnaphorsClusteringFormula.Template(a_1, a_2)$   
           $\rightarrow hasSameAntecedent(a_1, a_2)$
- 

Formula template for bridging antecedent selection

- $f_{r1}$      $\forall a \in A \forall e \in E : bridgingAnaResolutionFormula.Template1(a, e) \rightarrow isBridging(a, e)$   
 $f_{r2}$      $\forall a \in A \forall e \in E_a : bridgingAnaResolutionFormula.Template2(a, e) \rightarrow isBridging(a, e)$

bridging anaphor has at most one antecedent.<sup>23</sup>  $f_2$  models that a bridging anaphor should not appear before its antecedent.  $f_3$  and  $f_4$  model the reflexivity and transitivity of *sibling anaphor clustering*.  $f_5$  and  $f_6$  model that sibling anaphors share the same antecedent.

$f_c$  is the formula template for *sibling anaphor clustering*,  $f_{r1}$  and  $f_{r2}$  are the formula templates for *bridging antecedent selection*. Specific formulas instantiating  $f_c$  and  $f_{r1}/f_{r2}$  are described in Sections 5.2.1 and 5.2.2. In formulas instantiating  $f_{r2}$ , the set of antecedent candidates ( $E_a$ ) for bridging anaphor  $a$  is constructed on the basis of the anaphor's discourse scope (i.e., *local* or *non-local*) (described in Section 5.3).

## 5.2 Model: Features

We now describe all the features we use. The only additional manually annotated resource we need for feature extraction is WordNet.

*5.2.1 Features for Sibling Anaphor Clustering.* Table 11 shows the formulas for predicting sibling anaphors. Each formula is associated with a weight  $w$  learned during training. The polarity of the weights is indicated by the leading + or −.

$f_1$  captures that syntactically parallel anaphors are likely to be siblings. We define bridging anaphors  $a_1$  and  $a_2$  to be syntactically parallel if (i)  $a_1$  and  $a_2$  are coordinated NPs (e.g., **opposition** and **ruling-party members** in Example (24)); or (ii)  $a_1$  and  $a_2$  are both subjects/objects of verbs in conjoined clauses (e.g., **Three** and **two** in Example (25)).

(24) But this time it's hurting **opposition** as well as **ruling-party members**.

(25) Back in 1964, the FBI had *five black agents*. **Three** were chauffeurs for J. Edgar Hoover, and **two** cleaned his house.

Semantically related bridging anaphors are likely to be sibling anaphors.  $f_2$  models this via head match (see the two occurrences of **residents** in Example (26)).

(26) After being inspected, *buildings with substantial damage* were color-coded. Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

In  $f_3$ , we predict semantically related anaphors which do not share the same head word (such as **limited access** and **one last entry** in Example (26)), using WordNet-based similarity measures by Pedersen, Patwardhan, and Michelizzi (2004) in SVM<sup>light</sup>.<sup>24</sup>

*5.2.2 Features for Bridging Antecedent Selection.* Each formula for bridging antecedent selection (Table 12) is associated with a weight  $w$  learned during training. The polarity of the weights is indicated by leading + or −. For some formulas the final weight consists of a learned weight  $w$  multiplied by a score  $d$  (e.g., the inverse distance between

<sup>23</sup> We do not model that bridging anaphors have multiple antecedent entities (Example (7)).

<sup>24</sup> We could also use word embeddings as similarity measures. The focus of this article is not on similarity measures but on the joint optimization of antecedent selection. We therefore leave the investigation of different similarity measures to future work.

**Table 11**

Formulas for sibling anaphor clustering.  $a_1, a_2$  are bridging anaphors,  $A$  is the set of bridging anaphors in the document, and  $w$  the weight learned from the data for the specific formula.

---

**Formulas for sibling anaphors clustering**

---

f1	+	(w)	$\forall a_1, a_2 \in A \text{ ParallelAnas}(a_1, a_2) \rightarrow \text{hasSameAntecedent}(a_1, a_2)$
f2	+	(w)	$\forall a_1, a_2 \in A \text{ sameHead}(a_1, a_2) \rightarrow \text{hasSameAntecedent}(a_1, a_2)$
f3	+	(w)	$\forall a_1, a_2 \in A \text{ relatedTo}(a_1, a_2) \rightarrow \text{hasSameAntecedent}(a_1, a_2)$

---

antecedent and anaphor). In these cases, the final weight for a ground formula does not just depend on the respective formula, but also on the specific constants.

The numeric features  $f5, f7$ , and  $f11$  in Table 12 are normalized to between 0 and 1 among all antecedent candidates of one anaphor. Given a bridging anaphor  $a_i$ , its antecedent candidate set  $E_{a_i}$  ( $e_{ij} \in E_{a_i}$ ) and the numeric score  $S_{ik}$  for the pair  $\{a_i, e_{ik}\}$ , the normalized value of  $S_{ik}$  (i.e.,  $NormS_{ik}$ ) is calculated as:

$$NormS_{ik} = \frac{S_{ik} - \min_j S_{ij}}{\max_j S_{ij} - \min_j S_{ij}} \quad (27)$$

In contrast, the variants of these features (i.e.,  $f6, f8$ , and  $f12$ ) tell whether the score of an anaphor-antecedent candidate pair is the highest among all pairs for this anaphor.

*Frequent Bridging Relations* (Table 12:  $f1$ – $f4$ ).  $f1$ – $f4$  capture four bridging relations using the semantic classes of anaphor and antecedent, namely, the ones between role persons and GPEs (*USA - the president*), role persons and organizations (*the college - the principal*), relations (*She - the husband*) and times (*September - a year earlier*; Example (28)).

(28) Production of cars rose to 801,835 units in *September* from **a year earlier**.

For the first two bridging types we do not penalize antecedent candidates that are far away from the anaphor ( $f1$  and  $f2$ ). This is because in news it is common that a globally salient GPE or organization is introduced in the beginning, then later NPs denoting their roles are used as bridging anaphors throughout the document. For personal as well as temporal relations we prefer close antecedents by including the distance between antecedent and anaphor in the weights since these two bridging relations are local phenomena. These restrictions might be genre-specific.

*Semantic features: preposition pattern* (Table 12:  $f5$  and  $f6$ ). Corpus-based patterns capture semantic connectivity between a bridging anaphor and its antecedent. The “NP of NP” pattern (Poesio et al. 2004a) is useful for part-of and attribute-of relations (e.g., *windows of a room*) but cannot cover all bridging relations (such as *sanctions against a country*). We therefore generalize it to a *preposition pattern* to capture diverse semantic relations.

First, we extract the three most highly associated prepositions for each anaphor from Gigaword (Parker et al. 2011) and Tipster (Harman and Liberman 1993). This leads to, for example, the prepositions  $\{\textit{against}, \textit{on}, \textit{in}\}$  for the anaphor **sanctions**. Then for each anaphor-antecedent candidate pair, we query the corpora using their head

**Table 12**

Formulas for bridging antecedent selection.  $a$  is a bridging anaphor,  $A$  the set of bridging anaphors in the document,  $e$  the antecedent candidate,  $E_a$  the set of the antecedent candidates for  $a$  according to  $a$ 's discourse scope, and  $E$  the set of antecedent candidates in the document.

---

**Formulas for bridging antecedent selection**


---

## Semantic class features

- f1 + ( $w$ )  $\forall a \in A \forall e \in E : \text{hasSemanticClass}(a, \text{"gpeRolePerson"}) \wedge$   
 $\text{hasSemanticClass}(e, \text{"gpe"}) \wedge \text{hasPairDistance}(e, a, d) \wedge d > 0$   
 $\rightarrow \text{isBridging}(a, e)$
- f2 + ( $w$ )  $\forall a \in A \forall e \in E : \text{hasSemanticClass}(a, \text{"otherRolePerson"}) \wedge$   
 $\text{hasSemanticClass}(e, \text{"org"}) \wedge \text{hasPairDistance}(e, a, d) \wedge d > 0$   
 $\rightarrow \text{isBridging}(a, e)$
- f3 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E : \text{hasSemanticClass}(a, \text{"relativePerson"})$   
 $\wedge \text{hasSemanticClass}(e, \text{"person"} \star) \wedge \text{hasPairDistanceInverse}(e, a, d)$   
 $\rightarrow \text{isBridging}(a, e)$
- f4 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E : \text{hasSemanticClass}(a, \text{"date|time"})$   
 $\wedge \text{hasSemanticClass}(e, \text{"date|time"}) \wedge \text{hasPairDistanceInverse}(e, a, d)$   
 $\rightarrow \text{isBridging}(a, e)$

## Semantic features

- f5 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E_a : \text{relativeRankPrepPattern}(a, e, d) \rightarrow \text{isBridging}(a, e)$
- f6 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isTopRelativeRankPrepPattern}(a, e) \rightarrow \text{isBridging}(a, e)$
- f7 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E_a : \text{relativeRankVerbPattern}(a, e, d) \rightarrow \text{isBridging}(a, e)$
- f8 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isTopRelativeRankVerbPattern}(a, e) \rightarrow \text{isBridging}(a, e)$
- f9 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E_a : \text{isPartOf}(a, e) \wedge \text{hasPairDistanceInverse}(e, a, d)$   
 $\rightarrow \text{isBridging}(a, e)$

## Salience features

- f10 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{predictedGlobalAnte}(e) \wedge \text{hasPairDistance}(e, a, d)$   
 $\wedge d > 0 \rightarrow \text{isBridging}(a, e)$
- f11 + ( $w \cdot d$ )  $\forall a \in A \forall e \in E_a : \text{relativeRankDocSpan}(a, e, d) \rightarrow \text{isBridging}(a, e)$
- f12 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isTopRelativeRankDocSpan}(a, e) \rightarrow \text{isBridging}(a, e)$

## Lexical features

- f13 - ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isSameHead}(a, e) \rightarrow \text{isBridging}(a, e)$
- f14 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isPremodOverlap}(a, e) \rightarrow \text{isBridging}(a, e)$

## Syntactic features

- f15 - ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isCoArgument}(a, e) \rightarrow \text{isBridging}(a, e)$
- f16 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{synParallelStructure}(a, e) \rightarrow \text{isBridging}(a, e)$
- f17 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isClosestNominalModifier}(a, e) \rightarrow \text{isBridging}(a, e)$
- f18 + ( $w$ )  $\forall a \in A \forall e \in E_a : \text{isPredictSetBridging}(a, e) \rightarrow \text{isBridging}(a, e)$

words “*anaphor preposition antecedent*” (e.g. “sanction(s) against/on/in countr(y/ies)”). We replace proper names with fine-grained named entity types (using a gazetteer). Raw query hit counts are converted into Dunning root log-likelihood ratio scores<sup>25</sup> and then normalized using Equation (27). Table 13 shows some raw hit counts of the preposition pattern queries, the corresponding Dunning root log-likelihood ratio scores, and the normalized scores for the bridging anaphor **sanctions** and its antecedent candidates.

---

25 A variation of Dunning log-likelihood ratio (Dunning 1993) proposed by Dunning in [http://mail-archives.apache.org/mod\\_mbox/mahout-user/201001.mbox](http://mail-archives.apache.org/mod_mbox/mahout-user/201001.mbox).

**Table 13**  
An example of the preposition pattern feature. *RootLLR* = Dunning root log-likelihood ratio scores.

Anaphor	Antecedent Candidate	RawCount	RootLLR	NormalizedScore
sanctions	<i>the country</i>	6,817	81.44	1.00
sanctions	<i>apartheid</i>	26	4.8	0.32
sanctions	<i>further punishment</i>	9	-1.88	0.26
sanctions	...	...	...	...

*Semantic features: verb pattern* (Table 12: *f7* and *f8*). Set-membership relations between anaphor and antecedent evade the *preposition pattern*, because the anaphor often has no common noun head (Example (29)). However, in such a bridging relation, the antecedent is semantically compatible with the verb the anaphor depends on. In Example (29), *farmers travel* is more frequent than traveling pigs or dawns. We measure the compatibility between the antecedent candidates and the verb the anaphor depends on.

- (29) The cost of raising a pig kept bounding ahead of the return of selling one. *The farmers* stayed angry. [1 sent] At dawn on a cool day, **hundreds** travel to the private market in Radzymin [...]

Anaphors whose lexical head is an indefinite pronoun or a number are potential set bridging cases. We extract the verbs on which these potential set bridging anaphors depend (in our example, the verb *travel*). Finally, for each antecedent candidate, subject-verb, verb-object, or preposition-object queries<sup>26</sup> are executed against the Web 1T 5-gram corpus (Brants and Franz 2006). Raw hit counts are transformed into Dunning root log-likelihood ratio scores, then normalized as described in Equation (27).

*Semantic features: Part-of relation* (Table 12: *f9*). We use WordNet to decide whether a (possibly inherited) part-of relation holds between an anaphor and antecedent candidate.

*Salience features* (Table 12: *f10–f12*). Salient entities are preferred as bridging antecedents. In contrast to Poesio et al. (2004a), we find that bridging anaphors with distant antecedents are common if the antecedent is the global focus (Grosz and Sidner 1986).

*f10* models global salience by semantic connectivity to all bridging anaphors in the document. For each bridging anaphor  $a \in A$  and each entity  $e \in E$ , let  $score(a, e)$  be the preposition pattern score (*f5* in Table 12). We calculate the global semantic connectivity score  $e_{sal}$  for each  $e \in E$  as follows:  $e_{sal} = \sum_{a \in A} score(a, e)$ . If an entity appears in the headline<sup>27</sup> and also has the highest global semantic connectivity score among all entities in  $E$ , then this entity is predicted as globally salient for this document. Not every document has a globally salient entity.

26 The query form (i.e., subject-verb, verb-object, or preposition-object) is decided by the syntactic relation between the anaphor and its dependent verb/preposition.  
27 The texts in OntoNotes are not shown with headlines. However, the same texts are included in the Tipster corpus, from which we can extract the headlines.

$f_{11}$  and  $f_{12}$  capture salience by computing the span of text (measured in sentences) in which the antecedent candidate entity is mentioned divided by the number of sentences in the document.

*Lexical features* (Table 12:  $f_{13}$ – $f_{14}$ ). The *isSameHead* feature (Table 12:  $f_{13}$ ) checks whether antecedent candidates have the same head as the anaphor: This is rarely the case in bridging (except in some cases of set bridging and spatial/temporal sequence, see Example (30)) and can therefore be used to exclude antecedent candidates.

- (30) *His truck* is parked across the field, in a row of grain sellers. [2 sent.] The farmer at **the next truck** shouts, “Wheat!”

*isPremodOverlap* (Table 12:  $f_{14}$ ) determines the antecedent for compound noun anaphors whose head is preminally modified by the antecedent head (see Example (31)).

- (31) ... it doesn’t make *the equipment needed to produce those chips*. And IBM worries that the Japanese will take over **that equipment market**.

*Syntactic features: CoArgument* (Table 12:  $f_{15}$ ). The *CoArgument* feature excludes subjects from being antecedents for the object in the same clause, such as excluding “the Japanese” in Example (31) as antecedent for **that equipment market**.

*Syntactic features: intra-sentential syntactic parallelism* (Table 12:  $f_{16}$ ). If a noun phrase precedes a bridging anaphor in a different clause within the same sentence and both occupy the same syntactic role, it is likely that this noun phrase is the antecedent of the bridging anaphor. In Example (32), the anaphor and the antecedent are both objects of the verbs in the conjoined clauses. In Example (33), the anaphor and the antecedent both occupy the subject positions of the two conjoined clauses.

- (32) Poland must privatize *industry* and eliminate **subsidies** to stabilize its currency.
- (33) *One building* was upgraded to red status while people were taking things out, and **a resident who wasn’t allowed to go back inside** called up the stairs to his girlfriend ...

*Syntactic features: inter-sentential syntactic modification* (Table 12:  $f_{17}$ ). Laparra and Rigau’s (2013) work on implicit semantic role labeling assumes that different occurrences of the same predicate in a document likely maintain the same argument fillers. Therefore we can identify the antecedent of a bridging anaphor  $a$  by analyzing the nominal modifiers in other NPs with the same head word as  $a$ .<sup>28</sup> Whereas Laparra and Rigau’s work is restricted to ten predicates, we consider all bridging anaphors in ISNotes. In  $f_{17}$ , we predict antecedents for bridging anaphors by performing the following two steps:

1. For each bridging anaphor  $a$ , we take its head lemma  $a_h$  and collect all prenominal, possessive, and prepositional modifiers of other occurrences of  $a_h$  in the document. All realizations of these modifications that precede  $a$  form the antecedent candidate set  $Ante_a$ .

<sup>28</sup> Note that the bridging anaphor  $a$  is not coreferent to these other NPs with the same head word. Otherwise, its information status would be old and not bridging.



2. We choose the most recent mention from  $Ante_a$  as the predicted antecedent for the bridging anaphor  $a$ .

In Example (34), to resolve the bridging anaphor **heavy damage** to its antecedent *the quake, which registered 6.9 on the Richter scale*, we first check the other occurrences of the lemma “damage” and analyze their nominal modifiers—that is, one modifier is “area” (supported by *damage in the six - county San Francisco Bay area*) and the other modifier is “quake” (supported by *quake damage*). We then collect all mentions whose syntactic head is “area” or “quake” and that precede the anaphor in  $Ante_a$  (i.e., *the six-county San Francisco Bay area* and *the quake, which registered 6.9 on the Richter scale*). Finally, the most recent mention in  $Ante_a$  is predicted to be the antecedent.

- (34) Estimates of [damage in [the six - county San Francisco Bay area]] neared \$5 billion, excluding the cost of repairing the region’s transportation system.  
 ...  
 Part of the bridge collapsed in [*the quake, which registered 6.9 on the Richter scale*].  
 ... ..  
 While many of these buildings sustained **heavy damage**, little of that involved major structural damage.  
 ... ..  
 On Friday, during a visit to California to survey [*quake damage*], President Bush promised to “meet the federal government’s obligation” to assist relief effort.

*Syntactic features: hypertheme antecedent prediction for set sibling anaphors* (Table 12: f18). The VerbPattern features (Table 12: f7 and f8) only apply to a few typical set bridging cases, such as **None** in Example (10), here repeated as Example (35). Other set bridging anaphors (i.e., **One man** and **A food caterer**) are not covered by these features because they are not indefinite pronouns or numbers.

- (35) Still, [*employees*]<sub>hypertheme</sub> do occasionally try to smuggle out a gem or two.  
 [**One man**]<sub>theme</sub> wrapped several diamonds in the knot of his tie.  
 [**A food caterer**]<sub>theme</sub> stashed stones in the false bottom of a milk pail.  
 [**None**]<sub>theme</sub> made it past the body searches and X-rays of mine security.

Set bridging anaphors are often siblings (e.g., **One man**, **A food caterer**, and **None** are all elements of the set provided by *employees* in Example (35)). The information structure pattern we observe here is *Hypertheme–theme* (Daneš 1974). We predict heuristically the “themes” (set sibling anaphors) and their “Hypertheme” (antecedent). We first predict set sibling anaphors by expanding “typical” set bridging anaphors (e.g., **None** in Example 35) to their syntactically parallel neighbors (e.g., **One man** and **A food caterer**). We then predict the closest mention among all plural, subject mentions from the sentence immediately preceding the first anaphor as the antecedent for all (predicted) set sibling anaphors. If such a mention does not exist, the closest mention among all plural, object mentions from the sentence immediately preceding the first anaphor is predicted to be the antecedent. In Example (35), *employees* is predicted to be the antecedent for all (predicted) set sibling anaphors.

### 5.3 Method: Discourse Scope for Antecedent Candidate Selection

We use a new method, *d-scope-salience*, to form the list of antecedent candidates on the basis of the anaphor's discourse scope and apply it in the formulas  $f5$ – $f18$  in Table 12.<sup>29</sup>

*Motivation.* Ranking-based approaches for bridging antecedent selection need to tackle two interacting problems: (1) first, creating a list of antecedent candidates, (2) then, choosing an antecedent from this list. Once implausible candidates are removed from the list in (1), selecting the correct antecedent becomes an easier task in (2). Previous work (Markert, Nissim, and Modjeska 2003; Poesio et al. 2004a; Lassalle and Denis 2011) uses a static sentence window to construct the candidate list. However, this approach is problematic. If the window is too small, we miss too many correct antecedents. For example, 24% of anaphors in ISNotes would miss their antecedent if we used a two-sentence window (Section 3). If it is too large, we include too much noise in learning. In addition, whether more distant antecedents should be included might depend both on the salience properties of the antecedent and the place that the anaphor has in discourse.

We address this problem by proposing the *discourse scope* for an anaphor. Discourse entities have different scopes: Some contribute to the main topic and interact with distant entities (globally salient entities), and others focus on subtopics and only interact with nearby entities (locally salient entities). In Figure 4, the globally salient entity *Marina* in  $s1$  has a long forward lifespan, so that it can be accessed by both close and distant anaphors, **a resident** in  $s2$  and **residents** in  $s36$ . In contrast, the locally salient entity *buildings with substantial damage* in  $s24$  has a short forward lifespan, therefore it can only be accessed by nearby subsequent anaphors, **residents** and **limited access** in  $s25$ . Accordingly, anaphors that have non-local discourse scopes can access both locally and distant globally salient entities, whereas anaphors that have local discourse scopes can only access nearby locally salient entities. In consequence, we can add globally or locally salient entities to antecedent candidate lists for bridging anaphors according to their discourse scopes. The challenge is how to decide the discourse scopes for bridging anaphors automatically and how to model salience.

*Salience of Antecedents.* For each bridging anaphor  $a \in A$ , we define three antecedent candidate sets according to different salience measures:  $E_A^{\text{globalSal1}}$ ,  $E_A^{\text{globalSal2}}$ , and  $E_a^{\text{localSal}}$ :

- $E_A^{\text{globalSal1}}$  includes the top  $p$  percent salient entities in the text measured through the numbers of mentions in coreference chains.
- $E_A^{\text{globalSal2}}$  is the set of globally salient entities measured by the global semantic connectivity score (described in  $f10$  in Table 12). For each document, we create a list by ranking all entities according to their semantic connectivity to all anaphors. An entity is added to  $E_A^{\text{globalSal2}}$  if it ranks among the top  $k$  in this list and appears in the headline.

<sup>29</sup> Semantic class constraints ( $f1$ – $f4$  in Table 12) strongly indicate bridging. Hence, the antecedent access scope of an anaphor in these constraints is not strongly connected to the anaphor's discourse scope.


No DiscourseRel	non-local	s1: In the hard - hit <i>Marina</i> neighborhood, life after the earthquake is often all too real, but sometimes surreal.
		s2: Some scenes: -- Saturday morning, <b>a resident</b> was given 15 minutes to scurry into a sagging building and reclaim what she could of her life's possessions.
		...
Expansion.Restatement	local	s24: After being inspected, <i>buildings with substantial damage</i> were color - coded.
		s25: Green allowed <b>residents</b> to re-enter; yellow allowed <b>limited access</b> ; red allowed <b>residents one last entry</b> to gather everything they could within 15 minutes.
		...
Expansion.Conjunction	local	s34: <i>One building</i> was upgraded to red status while people were taking things out, and <b>a resident who wasn't allowed to go back inside</b> called <i>up the stairs</i> to his girl friend, telling her keep sending things down to <b>the lobby</b> .
		...
No DiscourseRel	non-local	s36: Enforcement of restricted - entry rules was sporadic, <b>residents</b> said.
Discourse Relation	anaLifeSpan	Bridging: 

Figure 4  
Global and local salience in bridging.

- The set  $E_a^{localSal}$  consists of locally salient entities. We approximate the entity’s local salience by the head position of its mention in the parse tree. Mentions preceding  $a$  in the same sentence and in the previous two sentences are added to  $E_a^{localSal}$  if the distance from their head to the root of the sentence’s dependency parse tree is less than threshold  $t$ .

*Anaphors’ discourse scopes.* We postulate that some discourse relations indicate the discourse scope of an anaphor. Here we use the discourse relation *Expansion* as defined in the Penn Discourse Treebank (Prasad et al. 2008). In this relation, the second argument elaborates on the first one and therefore most entities in the second argument contribute to local instead of global entity coherence. Therefore, we define two types of discourse scopes for bridging anaphora: *local* and *non-local*. If a bridging anaphor appears in argument 2 of an *Expansion* relation, it has *local* discourse scope; otherwise, it has *non-local* discourse scope.

*Antecedent candidate list for an anaphor via d-scope-salience.* We select the antecedent candidates for an anaphor via its discourse scope: For a *local* anaphor, only locally salient entities from the local window ( $E_a^{localSal}$ ) are allowed; for a *non-local* anaphor, apart from  $E_a^{localSal}$ , globally salient entities ( $E_A^{globalSal1}$  and  $E_A^{globalSal2}$ ) are also allowed.

5.4 Results and Discussion

We conduct experiments on the ISNotes corpus via 10-fold cross-validation on documents. We use the OntoNotes named entity and syntactic annotation as well as the Penn Discourse Treebank annotation for feature extraction. In each fold, we first choose ten documents randomly from the training set as the development set to estimate the values

of the parameters  $p$ ,  $k$ , and  $t$  in  $E_A^{\text{globalSal1}}$ ,  $E_A^{\text{globalSal2}}$ , and  $E_a^{\text{localSal}}$ , respectively,<sup>30</sup> then the whole training set is trained again using the optimized parameters.

**5.4.1 Mention-Entity Setting and Mention-Mention Setting.** In the mention-entity setting, entity information is based on the OntoNotes coreference annotation. We resolve bridging anaphors to entity antecedents. Features are extracted by using entity information. For instance, the semantic class of an entity is the majority semantic class of all its mention instantiations. The raw hit counts of the preposition pattern query for a bridging anaphor  $a$  and its antecedent candidate  $e$  ( $f5$  and  $f6$  in Table 12) is the maximum count among all instantiations of  $e$ . The distance between a bridging anaphor  $a$  and its antecedent candidate  $e$  is the distance between  $a$  and the closest mention instantiation of  $e$  preceding  $a$ .

In the mention-mention setting, we resolve bridging anaphors to mention antecedents and do not use any coreference information in the model or feature extraction. In this setting, we use “string match” for  $f11/f12$  in Table 12 and  $E_A^{\text{globalSal1}}$  in Section 5.3 to measure the salience of the mention antecedent candidates.

**5.4.2 Evaluation Metrics.** We measure accuracy on the number of bridging anaphors, instead of on all links between bridging anaphors and their antecedent instantiations. We calculate how many bridging anaphors are correctly resolved among all bridging anaphors. In the mention-entity setting, where the gold entity information is given, a bridging anaphor is counted as correctly resolved if the model links the anaphor to its entity antecedent. In the mention-mention setting, where the gold entity information is not given, a bridging anaphor is counted as correctly resolved if the model links the anaphor to one of its preceding antecedent instantiations. Statistical significance is measured using McNemar’s  $\chi^2$  test (McNemar 1947).

**5.4.3 Evaluation of Our New Local Features and Antecedent Candidate Selection.** To evaluate only the impact of our local features (Table 12) and the new antecedent candidate selection strategy (*d-scope-salience*, Section 5.3), we compare several pairwise machine learning models that successively build on each other. The pairwise model is widely used in coreference resolution (Soon, Ng, and Lim 2001) and has been used for bridging in Poesio et al. (2004a). Similar to the latter, we use it for bridging antecedent selection in the following way: Given an anaphor mention  $a$  and the set of antecedent candidate entities  $E_a$  that appear before  $a$ , we create a pairwise instance  $(a, e)$  for every  $e \in E_a$ . A binary decision whether  $a$  is bridged to  $e$  is made for each instance  $(a, e)$  separately. Finally, we explore the *best first* strategy (Ng and Cardie 2002) to choose one antecedent for each bridging anaphor. As we evaluate in the mention-entity setting, full coreference information is used in feature computation for all models.

*baseline1\_NB* and *baseline2\_NB*. We reimplement the algorithm from Poesio et al. (2004a) as a baseline. It is a pairwise naive Bayes classifier that classifies every anaphor-potential antecedent pair as true antecedent or not. We use the standard naive Bayes settings in WEKA (Witten and Frank 2005) with a *best first* strategy for choosing the correct antecedent (as described above).

<sup>30</sup> The parameter is estimated using a grid search over  $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ ,  $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ , and  $t \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ .

**Table 14**  
Feature set used in Poesio et al. (2004) and in our three baselines.

Group	Feature	Value
lexical	Google distance	numeric
	WordNet distance	numeric
salience	utterance distance	numeric
	local first mention	boolean
	global first mention	boolean

Because they did not explain whether they conducted the experiments under the mention-mention or the mention-entity setting, we assume they treated antecedents as entities. We use a two sentence (*baseline1\_NB*) and five sentence (*baseline2\_NB*) window for antecedent candidate selection.<sup>31</sup>

Poesio et al. (2004a) capture meronymy bridging relations via Google distance and WordNet distance (see Table 14). The former is the inverse value of Google hit counts for the *NP of NP* pattern query (e.g., *the windows of the center*). Because the Google API is no longer available, we use the Web 1T 5-gram corpus (Brants and Franz 2006) to extract the Google distance feature. We improve it by taking all information about entities via coreference into account and by replacing proper names with fine-grained named entity types (using a gazetteer). WordNet distance is the inverse value of the shortest path length between anaphor and antecedent candidate among all synset combinations. The other features measure the salience of an antecedent candidate. For instance, local first mention checks whether an antecedent candidate is realized in the first position of a sentence within the previous five sentences of the anaphor. Global first mention checks whether an antecedent candidate is realized in the first position of a sentence anywhere.

*baseline3\_SVM*. In *baseline3\_SVM*, we use the same features and the same antecedent candidate selection method as in *baseline1\_NB*, but replace naive Bayes with SVM<sup>light</sup>.<sup>32</sup> We stick with a two-sentence window as it performed on a par with the five-sentence window in the previous baselines.

*local1\_SVM*. In *local1\_SVM*, we use the same classifier and the same antecedent candidate selection method as in *baseline3\_SVM*, but replace the lexical features from Poesio et al. (2004a) (Table 14) with our *preposition pattern* features (*f5* and *f6* in Table 12).

*local2\_SVM*. On the basis of *local1\_SVM*, all other features from Table 12 (i.e., *f1–f4*, *f7–f18*) are added.

<sup>31</sup> Poesio et al. (2004a) used a five sentence window for antecedent candidate selection, because all antecedents in their corpus are within the previous five sentences of the anaphors.  
<sup>32</sup> We replace naive Bayes with SVM<sup>light</sup> because it can potentially deal better with imbalanced data. The SVM<sup>light</sup> parameter that handles data imbalance is set according to the ratio between positive and negative instances in the training set.

**Table 15**  
Results for bridging antecedent selection: Comparing pairwise models with different feature sets and antecedent candidate selection strategies. “Ante.” and “Acc.” stand for “Antecedent” and “Accuracy,” respectively. The **bolded** score indicates a significant improvement over all other models ( $p < 0.01$ ).

Model	Features	Ante. candidate list	Setting	Acc.
<i>baseline1_NB</i>	Poesio features	2-sentence-window	<i>mention-entity</i>	18.9
<i>baseline2_NB</i>	Poesio features	5-sentence-window	<i>mention-entity</i>	18.4
<i>baseline3_SVM</i>	Poesio features	2-sentence-window	<i>mention-entity</i>	19.8
<i>local1_SVM</i>	Poesio salience features + PrepPattern features ( <i>f</i> 5 and <i>f</i> 6 from Table 12)	2-sentence-window	<i>mention-entity</i>	29.1
<i>local2_SVM</i>	Poesio salience features + all features from Table 12	2-sentence-window	<i>mention-entity</i>	39.3
<i>local3_SVM</i>	Poesio salience features + all features from Table 12	<i>d-scope-salience</i>	<i>mention-entity</i>	<b>46.0</b>

*local3\_SVM*. On the basis of *local2\_SVM*, we apply our new method (*d-scope-salience*, Section 5.3) to select antecedent candidates for bridging anaphors.

*local1\_SVM* already outperforms the three baselines (*baseline1\_NB*, *baseline2\_NB*, and *baseline3\_SVM*) by about 10% (Table 15). This is due to normalizing the *preposition pattern* feature (Equation (27) in Section 5.2.2), and generalizing it (from the preposition *of* to appropriate prepositions for each anaphor) to capture more diverse semantic relations. This is important as our preposition pattern feature does not need more resources than the original Google distance feature in Poesio et al. (2004a) as it only depends on counts from unannotated corpora. The significant improvements of *local2\_SVM* indicate the contribution of our other features—however, these features sometimes need additional annotation in OntoNotes (such as the syntactic annotation) so the scenario is more idealized. Further improvements are achieved by *local3\_SVM*, which shows the positive impact of our advanced antecedent candidate selection strategy.

**5.4.4 Evaluation of the Joint Inference Model.** Simply porting our local model to MLNs (without including joint modeling and sibling anaphors clustering) does not improve performance (see Model *local\_MLN* in Table 16). The model *joint<sub>me</sub>* is the joint inference system described in Section 5.1 with all features for *sibling anaphors clustering* (Section 5.2.1) on top of all features for *bridging antecedent selection* (Section 5.2.2), using a mention-entity setting. We use *thebeast* to learn weights for the formulas and to perform inference.<sup>33</sup> *joint<sub>me</sub>* performs significantly better than the two local models (Table 16). This confirms our assumption that additional information from *sibling anaphors clustering* helps to resolve bridging anaphora.

<sup>33</sup> During training, we have 45 training instances in each fold. In fold0, the ground Markov network of *joint<sub>me</sub>* for the first training instance contains 2361 variables, and it takes around 3 minutes on an 8 CPU core machine to train the model.

**Table 16**  
Results for bridging antecedent selection: Comparing the local models to the joint inference model in different settings. The **bolded** score indicates a significant improvement over all other models ( $p < 0.01$ ).

	Setting	Model	Accuracy
local	mention-entity	local3_SVM	46.0
	mention-entity	local_MLN	46.4
joint inference	mention-entity	joint <sub>me</sub>	<b>50.7</b>
	mention-mention	joint <sub>mm</sub>	39.8
	mention-entity/mention	joint <sub>me:mm</sub>	44.2

The system *joint<sub>mm</sub>* includes the same features, sibling clustering, and antecedent selection as *joint<sub>me</sub>* but is trained and tested in the mention-mention setting. *joint<sub>me:mm</sub>* is trained in the mention-entity setting but tested in the mention-mention setting.

The performance of *joint<sub>mm</sub>* drops dramatically compared with *joint<sub>me</sub>* (Table 16). The representation of *sibling anaphors clustering* is noisier in this setting (e.g., two sibling anaphors may no longer share the same antecedent in the mention-mention setting) and features become weak. In Example (36), two sibling anaphors, **Employees** and **workers**, share the same entity antecedent represented by coreferent mentions (*Mobil Corp.*, *the company's*, and *Mobil*), but do not share the same mention antecedent in the mention-mention setting (e.g., *Mobil* is not the antecedent of the bridging anaphor **Employees**). Furthermore, knowing that *Mobil* is a company when using entity information (in the mention-entity setting) helps to resolve the bridging anaphor **workers**, whereas this information is not available in the mention-mention setting. In Example (36), the mention *the company's* is distant from the anaphor **workers**; therefore it is not included as an antecedent candidate for the anaphor **workers** in the mention-mention setting.

(36) Mobil Corp. is preparing to slash the size of its work force in the U.S., say individuals familiar with the company's strategy. **Employees** haven't yet been notified.  
...  
Some Mobil executives were dismayed that a reference to the cutbacks was included in the earning report before **workers** were notified.

*joint<sub>me:mm</sub>* performs significantly better than *joint<sub>mm</sub>*. Training the model in the mention-entity setting represents the phenomenon better than training in the noisy mention-mention setting.

**5.4.5 Error Analysis.** We conducted an error analysis for our best model *joint<sub>me</sub>*. First, anaphors with long distance antecedents are harder to resolve (see Table 17).

We now distinguish between sibling anaphors and non-sibling anaphors. The performance of *joint<sub>me</sub>* is 62.2% on sibling anaphors but only 34.8% on non-sibling anaphors. Global salience and links between related anaphors do indeed help to capture the behavior of sibling anaphors.

**Table 17**  
Antecedent selection accuracy with regard to anaphor-antecedent distance in *joint<sub>me</sub>*.

Sentence distance	#pairs	<i>joint<sub>me</sub></i>
0	175	59.4
1	260	46.9
2	90	50.0
≥3	158	44.3

The semantic knowledge we have is still insufficient. Typical problems are:

- (i) Cases with context-specific bridging relations. For example, in one text about the stealing of sago palms in California, we found the anaphor **the thieves** with the antecedent *palms*, which is not a very common semantic link.
- (ii) More frequently, we have cases where several good antecedents from a semantic perspective can be found. For example, two laws are discussed and the subsequent anaphor **the veto** could be the veto of either bills. Integration of the wider context apart from the two noun phrases in question is necessary in these cases. This can include the semantics of modification, whereas we currently consider only head nouns. Thus, the anaphor **the local council** would preferably be interpreted as *the council of a village* instead of *the council of a state* due to the occurrence of *local*.

Finally, 6% of the anaphors in our corpus have a non-NP antecedent. As we only extract NP phrases as potential candidate antecedents, we cannot handle these.

6. Unrestricted Bridging Resolution

Unrestricted bridging resolution recognizes bridging anaphors (beyond definite NPs only) and also finds links to antecedents (beyond meronymic relations only).

6.1 Method

We combine the two models from the previous two sections in a pipeline (Figure 5). Given extracted mentions, the system first predicts bridging anaphors by applying cascading collective classification (Section 4). It then predicts antecedents for these bridging anaphors (in the mention-mention setting) by applying joint inference trained in the mention-entity setting (Model *joint<sub>me\_mmm</sub>* in Section 5).<sup>34</sup>

6.2 Experiments and Results

We conduct experiments on ISNotes via 10-fold cross-validation on documents. We use an evaluation metric based on the number of bridging anaphors. The system predicts one unique antecedent for each predicted bridging anaphor. A link is counted as correct

<sup>34</sup> We use this model for antecedent selection for the pipeline model, as having full entity and coreference information in the test data is unrealistic.



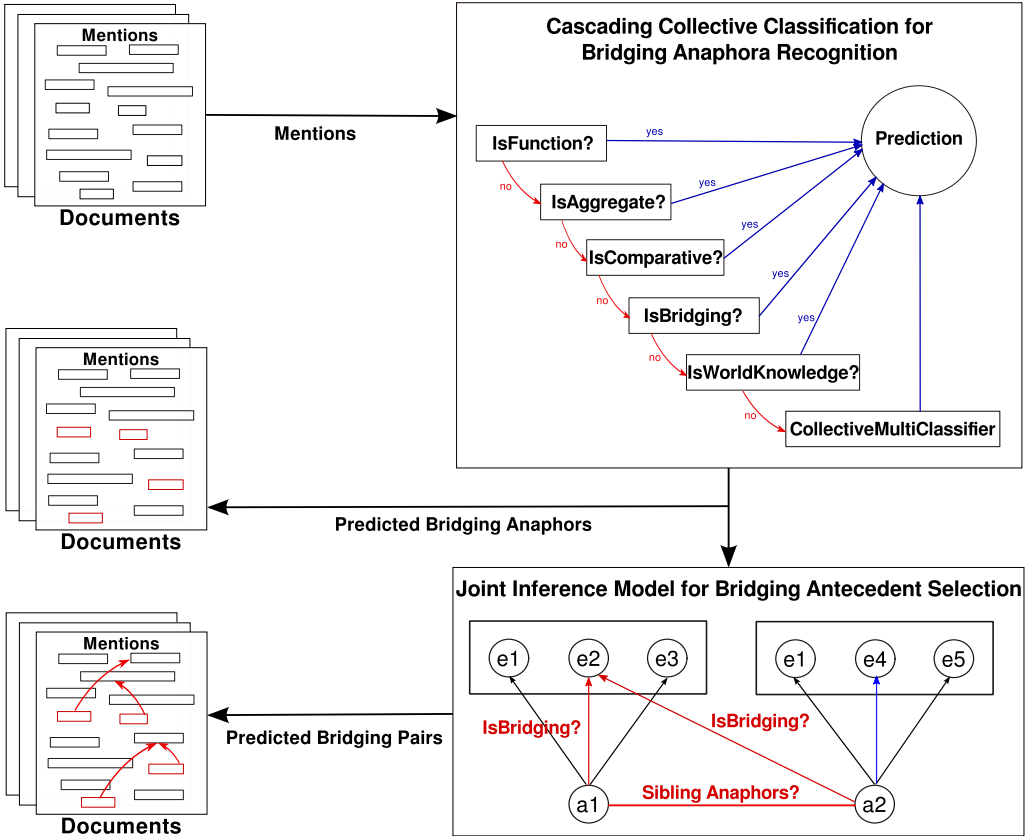


Figure 5  
The two-stage model for unrestricted bridging resolution.

if it recognizes the bridging anaphor correctly and links the anaphor to any instantiation of its antecedent preceding the anaphor. We use recall, precision, and F-score and the randomization test<sup>35</sup> on F-score for statistical significance.

6.2.1 *Baseline.* We compare our *pipeline model* to a learning-based model (*pairwise model*), adapted from the pairwise model widely used in coreference resolution (Soon, Ng, and Lim 2001).<sup>36</sup> In the *pairwise model* we first create an initial list of possible bridging anaphors  $A_{ml}$ , excluding as many obvious non-bridging mentions from the list as possible. A mention is added to  $A_{ml}$  if it (1) does not contain any other mentions, (2) is not modified by premodifications that strongly indicate comparative NPs, and (3) is not a pronoun or a proper name. Then for each NP  $a \in A_{ml}$ , a list of antecedent candidates  $C_a$  is created by including all mentions preceding  $a$  from the same as well as

<sup>35</sup> We use the package from <https://github.com/smartschat/art>.

<sup>36</sup> In Hou, Markert, and Strube (2014), we reimplement a previous rule-based system (Vieira and Poesio 2000) as the baseline. It suffers from a very low recall because it only considers meronymy bridging and compound noun anaphors whose head is prenominaly modified by the antecedent head. Therefore, we do not include it in this article as a baseline.

**Table 18**  
Experimental results for unrestricted bridging resolution. **Bolded** scores indicate significant improvements relative to other models ( $p < 0.01$ ).

	Bridging resolution		
	R	P	F
<i>pairwise model</i>	20.6	10.2	13.6
<i>pipeline model</i>	22.6	20.6	<b>21.6</b>

from the previous two sentences.<sup>37</sup> We create a pairwise instance  $(a, c)$  for every  $c \in C_a$ . In the decoding stage, the *best first* strategy (Ng and Cardie 2002) is used to predict bridging links. Specifically, for each  $a \in A_{ml}$ , we predict the bridging link to be the most confident pair  $(a, c_{ante})$  among all instances with the positive prediction. We provide this pairwise model with the same non-relational features as our two-stage model (Section 6.1); that is, features from Table 6 in Section 4.3.2 and Table 12 in Section 5.2.2. We use SVM<sup>light</sup> to conduct the experiments.<sup>38</sup>

*6.2.2 Results and Discussion.* Our *pipeline model* significantly outperforms the baseline (Table 18). Although the baseline models bridging anaphora recognition *and* antecedent selection together, it suffers from fewer positive training instances for each subtask because of its antecedent candidate selection strategy. In addition, we observe that diverse bridging relations in ISNotes, especially many context specific relations such as *pachinko* – **devotees** or *palms* – **the thieves**, lead to few training instances for each type of relation. As a result, generalizing is difficult for the learning-based approach. This is also the outcome of our earlier work (Hou, Markert, and Strube 2014), in which we propose a rule-based system for full bridging resolution on the same corpus. In this work, the rule-based system performs better than a learning-based approach (*pairwise model*) that has access to the same knowledge resources. Although the two-stage model outperforms our earlier rule-based system by 3.0 F-score points on bridging resolution, the result is still not satisfactory. This is due to the moderate performance in both stages. On bridging anaphora recognition, our best model (*CascadedCollective*) achieves an F-score of 46.1%. The errors in this stage are propagated to the second stage, where the accuracy of our best model (*joint<sub>me</sub>*) to choose antecedents for gold bridging anaphora is 50.7%. In future work, we would like to provide more training data to check whether the two-stage model benefits from it.

7. Conclusions

We presented the ISNotes corpus, which is annotated for a wide range of information status categories and full anaphoric information for the main anaphora types (i.e.,

<sup>37</sup> Initial experiments showed that increasing the window size more than two sentences decreases the performance.  
<sup>38</sup> To deal with data imbalance, the SVM<sup>light</sup> parameter is set according to the ratio between positive and negative instances in the training set.

coreference, bridging, and comparative). We developed a two-stage system for full bridging resolution, where bridging anaphors are not limited to definite NPs and bridging relations are not limited to meronymy. We proposed two joint inference models for information status recognition (including bridging recognition) and bridging antecedent selection, respectively. Our system achieves state-of-the-art performance or better for the three tasks (i.e., IS and bridging anaphora recognition, bridging antecedent selection, and bridging resolution) over reimplementations of previous approaches on ISNotes.

There are several open problems to be addressed. First, the results of our system might be improved with more annotations in the future. Given the difficulty of the task itself, we cannot expect that a large-scale corpus for bridging that is reliably annotated by linguists will appear any time soon. An option is to harvest potential bridging pairs by exploring semi-supervised or unsupervised learning approaches and combine these with expert/non-expert annotations. Second, our method should be tested in several other scenarios, such as its performance on other genres and its performance in less idealized conditions (such as automatically parsed corpora). Third, classifying bridging relations into fine-grained categories could be useful for other NLP applications, such as relation extraction across sentence boundaries and machine reading. Finally, bridging resolution, textual entailment, and implicit semantic role labeling are three standard tasks in NLP. They have some common properties and partially overlap. Recently, there are a few efforts that try to “bridge” boundaries between these tasks: Mirkin, Dagan, and Padó (2010) show that textual entailment recognition can benefit from bridging resolution; Stern and Dagan (2014) improve the performance of textual entailment recognition by exploring implicit semantic role labeling. It would be interesting to further explore the interactions between these three tasks, such as whether bridging anaphora recognition can benefit from the rich annotated data from FrameNet (e.g., *Null Instantiations*) or whether lexico-semantic resources widely used in textual entailment systems can be explored for bridging resolution.

## Acknowledgments

This work has been supported by the Research Training Group Coherence in Language Processing at Heidelberg University. Katja Markert received a Fellowship for Experienced Researchers by the Alexander-von-Humboldt Foundation. We thank HITS gGmbH for hosting Katja Markert and funding the annotation and the anonymous reviewers for their valuable feedback.

## References

- Abe, Naoki, Binanca Zadrozny, and John Langford. 2004. An iterative method for multi-class cost-sensitive learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3–11, Seattle, WA.
- Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.
- Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Asher, Nicholas and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15:83–113.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90, Montréal.
- Barzilay, Regina and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Björkelund, Anders, Kerstin Eckart, Arndt Riester, Nadja Schaufli, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the*

- 9th International Conference on Language Resources and Evaluation, pages 3222–3228, Reykjavik.
- Bos, Johan, Paul Buitelaar, and Anne Marie Mineur. 1995. Bridging as coercive accommodation. In *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95)*, pages 1–16, Edinburgh.
- Brants, Thorsten and Alex Franz. 2006. Web 1t 5-gram version 1. LDC2006T13, Philadelphia, PA, Linguistic Data Consortium.
- Burfoot, Clinton, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1506–1515, Portland, OR.
- Cahill, Aoife 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*, pages 817–825, Singapore.
- Cahill, Aoife and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the SIGdial 2012 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Seoul.
- Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Caselli, Tommaso and Irina Prodanof. 2006. Annotating bridging anaphors in Italian: In search of reliability. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1173–1176, Genoa.
- Cimiano, Philipp. 2006. Ingredients of a first-order account of bridging. In *Proceedings of the 5th International Workshop on Inference in Computational Semantics*, pages 139–144, Buxton.
- Clark, Herbert H. 1975. Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*, pages 169–174, Cambridge, MA.
- Clark, Herbert H. and Susan E. Haviland. 1977. Comprehension and the given-new contract. In Roy Freedle, editor, *Discourse Processes: Advances in Research and Theory*, volume 1. Ablex, Norwood, NJ, pages 1–40.
- Daneš, František. 1974. Functional sentence perspective and the organization of the text. In F. Daneš, editor, *Papers on Functional Sentence Perspective*, Prague: Academia, pages 106–128.
- Domingos, Pedro and Daniel Lowd. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Eckart, Kerstin, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics*, Springer, pages 65–76.
- Eckert, Miriam and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.
- Erkő, Feride and Jeanette K. Gundel. 1987. The pragmatics of indirect anaphors. In Jef Verschueren and Marcella Bertuccelli-Papi, editors, *The Pragmatic Perspective: Selected Papers from the 1985 International Pragmatics Conference*, Amsterdam, John Benjamins, pages 533–545.
- Fahrni, Angela and Michael Strube. 2012. Jointly disambiguating and clustering concepts and entities with Markov logic. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 815–832, Mumbai.
- Fraurud, Kari. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7:395–433.
- Gardent, Claire and Hélène Manuélian. 2005. Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.
- Garrod, Simon C. and Anthony J. Sanford. 1982. The mental representation of discourse in a focussed memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 1(1):21–41.
- Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 2000. Cognitive status and the form of indirect anaphors. *Verbum*, 22:79–102.
- Hahn, Udo, Michael Strube, and Katja Markert. 1996. Bridging textual ellipses. In *Proceedings of the 16th International Conference on Computational Linguistics*, volume 1, pages 496–501, Copenhagen.
- Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.
- Harman, Donna and Mark Liberman. 1993. TIPSTER Complete. LDC93T3A, Philadelphia, PA, Linguistic Data Consortium.
- Hawkins, John A. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Humanities Press, Atlantic Highlands, NJ.
- He, Haibo and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.
- Hirschman, Lynette and Nancy Chinchor. 1997. MUC-7 coreference task definition, <http://www.muc.saic.com/proceedings/>.
- Hobbs, Jerry R. 1978. Resolving pronominal references. *Lingua*, 44:311–338.
- Hobbs, Jerry R., Mark E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Hou, Yufang. 2016. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1880–1890, Osaka.
- Hou, Yufang, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, WA.
- Hou, Yufang, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, GA.
- Hou, Yufang, Katja Markert, and Michael Strube. 2014. A rule-based system for end-to-end bridging resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093, Doha.
- Irmer, Matthias. 2009. *Bridging Inferences in Discourse Interpretation*. Ph.D. thesis, Leipzig University.
- Jensen, David, Jennifer Neville, and Brian Gallagher. 2004. Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593–598, Seattle, WA.
- Joachims, Thorsten. 1999. Making large-scale support vector machine learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 169–184, MIT Press, Cambridge, MA.
- Kobayashi, Nozomi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, pages 1065–1074, Prague.
- Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA.
- Laparra, Egoitz and German Rigau. 2013. ImpAr: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1180–1189, Sofia.
- Lassalle, Emmanuel and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in French. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 35–46, Faro.
- LDC. 1993. Switchboard. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, PA.
- Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.
- Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen.

- Löbner, Sebastian. 1985. Definites. *Journal of Semantics*, 4:279–326.
- Löbner, Sebastian. 1998. Definite associative anaphora. Unpublished manuscript, Heinrich-Heine-Universität Düsseldorf.
- Macskassy, Sofus A. and Foster Provost. 2007. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.
- Markert, Katja, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island.
- Markert, Katja, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the Web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46, Budapest.
- Martschat, Sebastian and Michael Strube. 2014. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2070–2081, Doha.
- Matsui, Tomoko. 2000. *Bridging and Relevance*. John Benjamins.
- McNemar, Quinn. 1947. Note on the sampling errors of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 803–806, Lisbon.
- Mirkin, Shachar, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala.
- Mitchell, Alexis, Stephanie Strassel, Mark Przybicki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstain, Lisa Ferro, and Beth Sundheim. 2002. ACE-2 Version 1.0. LDC2003T11, Philadelphia, PA, Linguistic Data Consortium.
- Moschitti, Alessandro. 2006. Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–120, Trento.
- Nedoluzhko, Anna, Jiří Mirovský, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop at ACL-IJCNLP 2009*, pages 108–111, Suntec.
- Neville, Jennifer and David Jensen. 2003. Collective classification with relational dependency networks. In *Proceedings of the 2nd International Workshop on Multi-Relational Data Mining (MRDM-2003) at the International Conference on Knowledge Discovery and Data Mining*, pages 77–91, Washington, DC.
- Ng, Vincent and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, PA.
- Nissim, Malvina. 2006. Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 94–102, Sydney.
- Nissim, Malvina, 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*, pages 1023–1026, Lisbon.
- Omuya, Adinoyi, Vinodkumar Prabhakaran, and Owen Rambow. 2013. Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 802–807, Atlanta, GA.
- Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. LDC2011T07.
- Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity – Measuring the relatedness of concepts. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 267–270, Boston, MA.
- Poesio, Massimo. 2003. Associate descriptions and salience: A preliminary investigation. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pages 31–38, Budapest.

- Poesio, Massimo. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, MA.
- Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1220–1225, Las Palmas.
- Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004a. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 143–150, Barcelona.
- Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004b. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.
- Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Poesio, Massimo, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text*, pages 1–6, Madrid.
- Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech.
- Prince, Ellen F. 1981. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, Academic Press, New York, pages 223–255.
- Prince, Ellen F. 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*, John Benjamins, Amsterdam, pages 295–325.
- Rahman, Altaf 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, Edinburgh.
- Rahman, Altaf and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 798–807, Avignon.
- Reiter, Nils and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala.
- Riedel, Sebastian. 2008. Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 468–475, Helsinki.
- Riester, Arndt, 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*, pages 717–722, La Valetta.
- Rösiger, Ina and Simone Teufel. 2014. Resolving coreference and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55, Gothenburg.
- Ruppenhofer, Josef, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, pages 45–50, Uppsala.
- Sasano, Ryohei and Sadao Kurohashi. 2009. A probabilistic model for associative anaphora resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1455–1464, Singapore.
- Schwarz, Monika. 2000. *Indirekte Anaphern in Texten. Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Niemeyer, Tübingen, Germany.
- Somasundaran, Swapna, 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*, pages 170–179, Singapore.
- Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- Stern, Asher and Ido Dagan. 2014. Recognizing implied predicate-argument relationships in textual inference. In

- Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 739–744, Baltimore, MD.
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and Cambridge Computer Associates. 1966. *General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.
- Taskar, Ben, Pieter Abbeel, and Daphne Koller. 2002. Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 485–492, Edmonton.
- Taskar, Ben, Eran Segal, and Daphne Koller. 2001. Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 870–876, Seattle, WA.
- Vieira, Renata. 1998. *Definite Description Processing in Unrestricted Text*. Ph.D. thesis, University of Edinburgh.
- Vieira, Renata and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.
- Vieira, Renata and Simone Teufel. 1997. Towards resolution of bridging descriptions. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 522–524, Madrid.
- Voorhees, Ellen M. 2001. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtrieval Conference*, Gaithersburg, MD.
- Schulte im Walde, Sabine. 1998. Resolving bridging descriptions in high-dimensional space. Master's thesis, University of Edinburgh, Centre for Cognitive Science.
- Wang, Shuo and Xin Yao. 2012. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(4):1119–1130.
- Weischedel, Ralph, 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, PA, Linguistic Data Consortium.
- Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition, Morgan Kaufmann, San Francisco, CA.
- Zhou, Zhihua and Xuying Liu. 2010. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257.