Global Inference for Bridging Anaphora Resolution

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

Abstract

We present the first work on antecedent selection for bridging resolution without restrictions on anaphor or relation types. Our model integrates global constraints on top of a rich local feature set in the framework of Markov logic networks. The global model improves over the local one and both strongly outperform a reimplementation of prior work.

1 Introduction

Identity coreference is a relatively well understood and well-studied instance of entity coherence. However, entity coherence can rely on more complex, lexico-semantic, frame or encyclopedic relations than identity. Anaphora linking distinct entities or events this way are called *bridging* or *associative anaphora* and have been widely discussed in the linguistic literature (Clark, 1975; Prince, 1981; Gundel et al., 1993).[1] In Example 1, the phrases *the windows*, *the carpets* and *walls* can be felicitously used because they are semantically related via a part-of relation to their antecedent *the Polish center*.[2]

(1) ...as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. ...**The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

[1]Poesio and Vieira (1998) include cases where antecedent and anaphor are coreferent but do not share the same head noun. We restrict *bridging* to non-coreferential cases. We also exclude *comparative anaphora* (Modjeska et al., 2003)

[2]Examples are from OntoNotes (Weischedel et al., 2011). Bridging anaphora are typed in boldface; antecedents in italics.

Bridging is frequent amounting to between 5% (Gardent and Manuélian, 2005) and 20% (Caselli and Prodanof, 2006) of definite descriptions (both studies limited to NPs starting with *the* or non-English equivalents). Bridging resolution is needed to fill gaps in entity grids based on coreference only (Barzilay and Lapata, 2008). Example 1 does not exhibit any coreferential entity coherence. Coherence can only be established when the bridging anaphora are resolved. Bridging resolution may also be important for textual entailment (Mirkin et al., 2010).

Bridging resolution can be divided into two tasks, recognizing that a bridging anaphor is present and finding the correct antecedent among a list of candidates. These two tasks have frequently been handled in a pipeline with most research concentrating on antecedent selection only. We also handle only the task of antecedent selection.

Previous work on antecedent selection for bridging anaphora is restricted. It makes strong untested assumptions about bridging anaphora types or relations, limiting it to definite NPs (Poesio and Vieira, 1998; Poesio et al., 2004; Lassalle and Denis, 2011) or to part-of relations between anaphor and antecedent (Poesio et al., 2004; Markert et al., 2003; Lassalle and Denis, 2011). We break new ground by considering all relations and anaphora/antecedent types and show that the variety of bridging anaphora is much higher than reported previously.

Following work on coreference resolution, we apply a *local* pairwise model (Soon et al., 2001) for antecedent selection. We then develop novel semantic, syntactic and salience features for this task, showing strong improvements over one of the best known

Proceedings of NAACL-HLT 2013, pages 907–917,

prior models (Poesio et al., 2004).

However, this local model classifies each anaphor-antecedent candidate pair in isolation. Thus, it neglects that bridging anaphora referring to a single antecedent often occur in clusters (see Example 1). It also neglects that once an entity is an antecedent for a bridging anaphor it is more likely to be used again as antecedent. In addition, such local models construct the list of possible antecedent candidates normally relying on a window size constraint to restrict the set of candidates: is the window too small, we miss too many correct antecedents; is it too large, we include so many incorrect antecedents as to lead to severe data imbalance in learning.

To remedy these flaws we change to a *global* Markov logic model that allows us to:

- model constraints that certain anaphora are likely to share the same antecedent;
- model the global semantic connectivity of a salient potential antecedent to all anaphora in a text;
- consider the union of potential antecedents for all anaphora instead of a static window-sized constraint.

We show that this global model with the same local features but enhanced with global constraints improves significantly over the local model.

2 Related Work

Prior corpus-linguistic studies on bridging are beset by three main problems. First, reliability is not measured or low (Fraurud, 1990; Poesio, 2003; Gardent and Manuélian, 2005; Riester et al., 2010).[3] Second, annotated corpora are small (Poesio et al., 2004; Korzen and Buch-Kromann, 2011). Third, they are often based on strong untested assumptions about bridging anaphora types, antecedent types or relations, such as limiting it to definite NP anaphora (Poesio and Vieira, 1998; Poesio et al., 2004; Gardent and Manuélian, 2005; Caselli and Prodanof, 2006; Riester et al., 2010; Lassalle and Denis, 2011), to NP antecedents (all prior work) or to part-of relations between anaphor and antecedent (Markert et al., 2003; Poesio et al., 2004). In our own work (Markert et al., 2012) we established a corpus that circumvents these problems, i.e. human bridging recognition was reliable, it contains a medium number of bridging cases that allows generalisable statistics and we did not limit bridging anaphora or antecedents according to their syntactic type or relations between them. However, we only discussed human agreement on bridging recognition in Markert et al. (2012), disregarding antecedent annotation. We also did not discuss the different types of bridging in the corpus. We will remedy this in Section 3.

Automatic work on bridging distinguishes between recognition (Vieira and Poesio, 2000; Rahman and Ng, 2012; Cahill and Riester, 2012; Markert et al., 2012) and antecedent selection. Work on antecedent selection suffers from focusing on subproblems, e.g. only part-of bridging (Poesio et al., 2004; Markert et al., 2003) or definite NP anaphora (Lassalle and Denis, 2011). Most relevant for us is Lassalle and Denis (2011) who restrict anaphora to definite descriptions but have no other restrictions on relations or antecedent NPs (in a French corpus) with an accuracy of 23%. Also the evaluation setup is sometimes not clear: The high results in Poesio et al. (2004) cannot be used for comparison as they test unrealistically: they distinguish only between the correct antecedent and *one* or *three* false candidates (baseline of 50% for the former). They also restrict the phenomenon to part-of relations.

There is a partial overlap between bridging and implicit noun roles (Ruppenhofer et al., 2010). However, work on implicit noun roles is mostly focused on few predicates (e.g. Gerber and Chai (2012)). We consider all bridging anaphors in running text. The closest work to ours interpreting implicit role filling as anaphora resolution is Silberer and Frank (2012).

3 Corpus for Bridging: An Overview

We use the dataset we created in Markert et al. (2012) with almost 11,000 NPs annotated for information status including 663 bridging NPs and their antecedents in 50 texts taken from the WSJ portion of the OntoNotes corpus (Weischedel et al., 2011). Bridging anaphora can be any noun phrase. They

[3] Although the overall information status scheme in Riester et al. (2010) achieved high agreement, their confusion matrix shows that the anaphoric bridging category (BRI) is frequently confused with other categories so that the two annotators agreed on only less than a third of bridging anaphors.

are not limited to definite NPs as in previous work. In contrast to Nissim et al. (2004), antecedents are annotated and can be noun phrases, verb phrases or even clauses. Our bridging annotation is also not limited with regards to semantic relations between anaphor and antecedent.

In Markert et al. (2012) we achieved high agreement for the overall information status annotation scheme between three annotators (κ between 75 and 80, dependent on annotator pairs) as well as for all subcategories, including bridging (κ over 60 for all annotator pairings, over 70 for two expert annotators). Here, we add the following new results:

- Agreement for selecting bridging antecedents was around 80% for all annotator pairings.
- Surprisingly, only 255 of the 663 (38%) bridging anaphors are definite NPs, which calls into question the strategy of prior approaches to limit themselves to these types of bridging.
- NPs are the most frequent antecedents by far with only 42 of 663 (6%) bridging anaphora having a non-NP antecedent (mostly verb phrases).
- Bridging is a relatively local phenomenon with 71% of NP antecedents occurring in the same or up to 2 sentences prior to the anaphor. However, farther away antecedents are common when the antecedent is the global focus of a document.
- The semantic relations between anaphor and antecedent are extremely diverse with only 92 of 663 (14%) anaphors having a part-of/attribute-of antecedent (see Example 1) and only 45 (7%) anaphors standing in a set relationship to the antecedent (see Example 2). This contrasts with Gardent and Manuélian's (2005) finding that 52% of bridging cases had meronymic relations. We find many different types of relations in our corpus, including encyclopedic relations such as *restaurant — the waiter* as well as, frequently, relational person nouns as bridging anaphors such as *friend, husband, president*.
- There are only a few cases of bridging where surface cues may indicate the antecedent. First, some bridging anaphors are modified by a small number of adjectives that have more than one role filler, with the bridging relation often being temporal or spatial sequence between two entities of the same semantic type as in Example 3 (see also Lassalle and Denis (2011) for a discussion of such cases). Second, some anaphors are compounds where the nominal premodifier matches the antecedent head as in Example 4.

(2) Still *employees* do occasionally try to smuggle out a gem or two. **One man** wrapped several diamonds in the knot of his tie. **Another** poked a hole in the heel of his shoe. **None** made it past the body searches ...

(3) *His truck* is parked across the field ... The farmer at **the next truck** shouts ...

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

4 Models for Bridging Resolution

4.1 Pairwise mention-entity model

The pairwise model is widely used in coreference resolution (Soon et al., 2001). We adapt it for bridging resolution[4]: Given an anaphor mention m and the set of antecedent candidate entities E_m which appear before m, we create a pairwise instance (m, e) for every $e \in E_m$. A binary decision whether m is bridged to e is made for each instance (m, e) separately. A post-processing step to choose one antecedent is necessary (closest first or best first are common strategies). This model causes three problems for bridging resolution: First, the ratio between positive and negative instances is 1 to 17 even if only antecedent candidates from the current and the immediately preceding two sentences are considered. The ratio will be even worse with a larger window size. Therefore, usually a fixed window size is used restricting the set of candidates. This, however, causes a second problem: antecedents which are beyond the window cannot be found. In our data, only 81% of NP antecedents appear within the previous 5 sentences, and only 71% of NP antecedents appear within the previous 2 sentences. The third problem is a shortcoming of the pairwise model itself: decisions are made for each instance separately, ignoring

[4] Different from coreference, we treat an anaphor as a mention and an antecedent as an entity. The anaphor is the first mention of the corresponding entity in the document.

relations between instances. We resolve these problems by employing a global model based on Markov logic networks.

4.2 Markov Logic Networks

Bridging can be considered a document global phenomenon, where globally salient entities are preferred as antecedents and two or more anaphors having the same antecedent should be related or similar. Motivated by this observation, we explore Markov logic networks (Domingos and Lowd, 2009, MLNs) to model bridging resolution on the global discourse level.

MLNs are a powerful representation for joint inference with uncertainty. An MLN consists of a set of pairs (F_i, w_i), where F_i is a formula in first-order logic and w_i is its associated real numbered weight. It can be viewed as a template for constructing Markov networks. Given different sets of constants, an MLN will produce different ground Markov networks which may vary in size but have the same structure and parameters. For a ground Markov network, the probability distribution over possible worlds x is given by

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right) \qquad (1)$$

where $n_i(x)$ is the number of true groundings of F_i in x. The normalization factor Z is the partition function.

MLNs have been applied to many NLP tasks and achieved good performance by leveraging rich relations among objects (Poon and Domingos, 2008; Meza-Ruiz and Riedel, 2009; Fahrni and Strube, 2012, inter alia). We use *thebeast*[5] to learn weights for the formulas and to perform inference. *thebeast* employs cutting plane inference (Riedel, 2008) to improve the accuracy and efficiency of MAP inference for Markov logic.

With MLNs, we model bridging resolution globally on the discourse level: given the set M of all anaphors and sets of local antecedent candidates E_m for each anaphor $m \in M$, we select antecedents for all anaphors from $E = \bigcup_{m \in M} E_m$ at the same time. Table 1 shows the hidden predicates and formulas used. Each formula is associated with a weight. The polarity of the weights is indicated by the leading $+$ or $-$. The weight value (except for hard constraints) is learned from training data. For some formulas the final weight consists of a learned weight w multiplied by a score d (e.g. inverse distance between antecedent and anaphor). In these cases the final weight for a formula in a ground Markov network does not just depend on the respective formula, but also on the specific constants. We indicate such combined weights by the term $w \cdot d$.

We tackle the previously mentioned problems of the pairwise model: (1) We construct hard constraints to specify that each anaphor has at most one antecedent entity (Table 1: f1) and that the antecedent must precede the anaphor (f2). This eliminates the need for the post-processing step in the pairwise model. (2) We select the antecedent entity for each anaphor from the antecedent candidate entities pool E which alleviates the missing true antecedent problem in the pairwise model. Based on (1) and (2), MLNs allow us to express relations between anaphor-anaphor and anaphor-antecedent pairs ((m,n) or (m,e)) on the global discourse level improving accuracy by performing joint inference.

5 Features

5.1 Local features

5.1.1 Poesio et al.'s feature set

Table 2 shows the feature set proposed by Poesio et al. (2004) for part-of bridging. Google distance is the inverse value of Google hit counts for the *ofPattern* query (e.g. *the windows of the center*). WordNet distance is the inverse value of the shortest path length between an anaphor and an antecedent candidate among all synset combinations. These features are supposed to capture the meronymy relation between anaphor and antecedent. The other ones measure the salience of the antecedent candidate.

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

Table 2: Poesio et al.'s feature set

[5] http://code.google.com/p/thebeast

Hidden predicates			
p1			$isBridging(m,e)$
p2			$hasSameAntecedent(m,n)$
Formulas			
Hard constraints			
f1			$\forall m \in M: \ \lvert e \in E : isBridging(m,e) \rvert \leq 1$
f2			$\forall m \in M \forall e \in E: \ hasPairDistance(e,m,d) \wedge d < 0 \rightarrow \neg isBridging(m,e)$
f3			$\forall m,n \in M: \ m \neq n \wedge hasSameAntecedent(m,n)$ $\rightarrow hasSameAntecedent(n,m)$
f4			$\forall m,n,l \in M: \ m \neq n \wedge m \neq l \wedge n \neq l \wedge hasSameAntecedent(m,n)$ $\wedge hasSameAntecedent(n,l) \rightarrow hasSameAntecedent(m,l)$
f5			$\forall m,n \in M \forall e \in E: \ m \neq n \wedge hasSameAntecedent(m,n) \wedge isBridging(m,e)$ $\rightarrow isBridging(n,e)$
f6			$\forall m,n \in M \forall e \in E: \ m \neq n \wedge isBridging(m,e) \wedge isBridging(n,e)$ $\rightarrow hasSameAntecedent(m,n)$
Discourse level formulas			
f7	$+$	(w)	$\forall m \in M \forall e \in E: \ predictedGlobalAnte(e) \wedge hasPairDistance(e,m,d)$ $\wedge d > 0 \rightarrow isBridging(m,e)$
f8	$+$	(w)	$\forall m,n \in M \ conjunction(m,n) \rightarrow hasSameAntecedent(m,n)$
f9	$+$	(w)	$\forall m,n \in M \ sameHead(m,n) \rightarrow hasSameAntecedent(m,n)$
f10	$+$	(w)	$\forall m,n \in M \ similarTo(m,n) \rightarrow hasSameAntecedent(m,n)$
f11	$+$	(w)	$\forall m \in M \forall e \in E: \ hasSemanticClass(m, "rolePerson")$ $\wedge hasSemanticClass(e, "org\lvert gpe") \wedge hasPairDistance(e,m,d) \wedge d > 0$ $\rightarrow isBridging(m,e)$
f12	$+$	$(w \cdot d)$	$\forall m \in M \forall e \in E: \ hasSemanticClass(m, "relativePerson")$ $\wedge hasSemanticClass(e, "otherPerson") \wedge hasPairDistanceInverse(e,m,d)$ $\rightarrow isBridging(m,e)$
f13	$+$	$(w \cdot d)$	$\forall m \in M \forall e \in E: \ hasSemanticClass(m, "date")$ $\wedge hasSemanticClass(e, "date") \wedge hasPairDistanceInverse(e,m,d)$ $\rightarrow isBridging(m,e)$
Local formulas			
f14	$+$	(w)	$\forall m \in M \ \forall e \in E_m: \ isTopRelativeRankPrepPattern(m,e) \rightarrow isBridging(m,e)$
f15	$+$	(w)	$\forall m \in M \ \forall e \in E_m: \ isTopRelativeRankVerbPattern(m,e) \rightarrow isBridging(m,e)$
f16	$+$	$(w \cdot d)$	$\forall m \in M \ \forall e \in E_m: \ isPartOf(m,e) \wedge hasPairDistanceInverse(e,m,d)$ $\rightarrow isBridging(m,e)$
f17	$+$	(w)	$\forall m \in M \ \forall e \in E_m: \ isTopRelativeRankDocSpan(m,e) \rightarrow isBridging(m,e)$
f18	$-$	(w)	$\forall m \in M \ \forall e \in E_m: \ isSameHead(m,e) \rightarrow isBridging(m,e)$
f19	$+$	(w)	$\forall m \in M \ \forall e \in E_m: \ isPremodOverlap(m,e) \rightarrow isBridging(m,e)$
f20	$-$	(w)	$\forall m \in M \ \forall e \in E_m: \ isCoArgument(m,e) \rightarrow isBridging(m,e)$

Table 1: Hidden predicates and formulas used for bridging resolution (m, n, l represent mentions, M the set of bridging anaphora mentions in the whole document, e the antecedent candidate entity, E_m the set of local antecedent candidate entities for m, and $E = \bigcup_{m \in M} E_m$)

5.1.2 Other features

Since Poesio et al. (2004) deal exclusively with meronymy bridging, we have to extend the feature set to capture more diverse relations between anaphor and antecedent. All numeric features in Table 3 are normalized among all antecedent candidates of one anaphor. For anaphor m_i and its antecedent candidates E_{m_i} ($e_{ij} \in E_{m_i}$), the numeric score for pair $\{m_i,\ e_{ik}\}$ is S_{ik}. Then the value $NormS_{ik}$ for this pair is normalized (set to values between 0 and 1) as below:

$$NormS_{ik} = \frac{S_{ik} - \min_j S_{ij}}{\max_j S_{ij} - \min_j S_{ij}} \quad (2)$$

A second variant of numeric features tells whether the score of an anaphor-antecedent candidate pair is the highest among all pairs for this anaphor.

Group	Feature	Value
semantic	$_{feat1}$ preposition pattern	numeric
	$_{feat2}$ verb pattern	numeric
	$_{feat3}$ WordNet partOf	boolean
	$_{feat4}$ semantic class	nominal
salience	$_{feat5}$ document span	numeric
surface	$_{feat6}$ isSameHead	boolean
	$_{feat7}$ isPremodOverlap	boolean
syntactic	$_{feat8}$ isCoArgument	boolean

Table 3: Local features we developed

Preposition pattern ($feat1$). The *ofPattern* proposed by Poesio et al. (2004) is useful for part-of and attribute-of relations but cannot cover all bridging relations (such as *sanctions against a country*). We extend the *ofPattern* to a generalised *preposition pattern* by using the Gigaword (Parker et al., 2011) and the Tipster (Harman and Liberman, 1993) corpora (both automatically POS tagged and NP chunked for improving query match precision).

First, we extract the three most highly associated prepositions for each anaphor. Then for each anaphor-antecedent candidate pair, we use their head words to create the query *"anaphor preposition antecedent"*. To improve recall, we take lowercase, uppercase, singular and plural forms of the head word into account, and replace proper names by fine-grained named entity types (using a gazetteer). All raw hit counts are converted into the Dunning Root Loglikelihood association measure,[6] then normalized using Formula 2 within all antecedent candidates of one anaphor.

Verb pattern ($feat2$). A set-membership relation between anaphor and antecedent is often hard to capture by the *preposition pattern* because the anaphor often has no common noun head (see Example 2 in Section 3). Hence, we measure the compatibility of the antecedent candidates with the verb the anaphor depends on.

First, we hypothesise that anaphors whose lexical head is a pronoun or a number are potential set bridging cases and then extract the verb the anaphor depends on. In example 2, for the set anaphor **Another**, *poked* is the verb. Then for each antecedent candidate, subject-verb or verb-object queries are applied to the Web 1T 5-gram corpus (Brants and Franz, 2006). In this case, *employees poked* and *diamonds poked* are example queries. The hit counts are transformed into PMI and all pairs for one anaphor are normalized as described in Formula 2.

WordNet partOf relation ($feat3$). To capture part-of bridging, we extract whether the anaphor is part of the antecedent candidate in WordNet. To improve recall, we use hyponym information of the antecedent. If an antecedent e is a hypernym of x and an anaphor m is a meronym of x, then m is a meronym of e.

Semantic class ($feat4$). The anaphor and the antecedent candidate are assigned one of 16 coarse-grained semantic classes, e.g. location, organization, GPE, roleperson, relativePerson, otherPerson[7], product, language, NORP (nationalities, religious or political groups) and several classes for numbers (such as date, money or percent).

Salience feature ($feat5$). Salient entities are preferred as antecedents. We capture salience superficially by computing the *"antecedent document span"* of an antecedent candidate. We compute the

[6] `http://tdunning.blogspot.de/2008/03/surprise-and-coincidence.html`

[7] We use WordNet to extract lists for *rolePerson* (persons like *president* or *teacher* playing a role in an organization) and *relativePerson* (persons like *father* or *son* indicating that they have a relation with another person). Persons not in these two lists are counted as *otherPerson*.

span of text (measured in sentences) in which the antecedent candidate entity is mentioned. This is divided by the number of sentences in the whole document. This score is normalized using Formula 2 for all antecedent candidates of one anaphor.

Surface features ($feat6$-$feat7$). *isSameHead* ($feat6$) checks whether antecedent candidates have the same head as the anaphor: this is rarely the case in bridging anaphora (except in some cases of set bridging and spatial/temporal sequence, see Example 3) and can therefore be used to exclude antecedent candidates. *isPremodOverlap* ($feat7$) determines the antecedent for compound noun anaphors whose head is prenominally modified by the antecedent head (see Example 4).

Syntactic feature ($feat8$) The *isCoArgument* feature is based on the intuition that the subject cannot be the bridging antecedent of the object in the same clause. This feature excludes (some) close antecedent candidates. In Example 4, the antecedent candidate *the Japanese* isCoArgument with the anaphor **that equipment market**.

5.2 Global features for MLNs

$f1$-$f13$ in Table 1 are discourse level constraints. All antecedent candidates come from the antecedent candidates pool E in the whole document.

Global salience (Table 1: $f3$-$f10$). The salience feature in the pairwise model only measures the salience for candidates within the local window. However, globally salient antecedents are preferred even if they are far away from the anaphor. We model this from two perspectives:

$f7$ models the preference for globally salient antecedents, which we derive for each document. For $m \in M$ and $e \in E$, let $score(m, e)$ be the preposition pattern score for pair (m,e). Calculate pattern semantic salience score e_{sal} for each $e \in E$ as

$$e_{sal} = \sum_{m \in M} score(m, e) \tag{3}$$

If e appears in the title and also has the highest pattern semantic salience score e_{sal} among all e in E, then e is the predicted globally salient antecedent for this document. Note that global salience here is based on semantic connectivity to all anaphors in the document and that not every document has a globally salient antecedent.

$f3$-$f6$ and $f8$-$f10$ model that similar or related anaphors in one document are likely to have the same antecedent. To make the ground Markov network more sparse for more efficient inference, we add the hidden predicate ($p2$) and hard constraints ($f3$-$f6$) specifying relations among similar/related anaphors m, n and l (reflexivity and transitivity). Formulas $f8$-$f10$ explore three different ways (syntactic and semantic) to compute the similarity between two anaphors. In $f10$, we use SVM^{light} (similarity scores from WordNet plus sentence distance as features) to predict whether two anaphors not sharing the same head are similar or not.

Frequent bridging relations (Table 1: $f11$-$f13$). Three common bridging relations are restricted by semantic class of anaphor and antecedent (see also Section 3). It is worth noting that in formula $f11$ (modeling that a role person mention like *president* or *chairman* prefers organization or GPE antecedents), we do not penalize the antecedents far away from the anaphor. In formula $f12$ (modeling that a relativePerson mention such as *mother* or *husband* prefers close person antecedents) and $f13$, we prefer close antecedents by including the distance between antecedent and anaphor into the weights.

MLN formulation of local features (Table 1: $f14$-$f20$). Corresponding to features of the pairwise model (Table 3) – we exclude only semantic class as this is modelled globally via features $f11$-$f13$. These local features are only used for an anaphor m and its local antecedent candidate e from E_m.

6 Experiments and Results

6.1 Experimental setup

We perform experiments on our gold standard corpus via 10-fold cross-validation on documents. We use gold standard mentions, true coreference information, and the OntoNotes named entity and syntactic annotation layers for feature extraction.

6.2 Improved baseline

We reimplement the algorithm from Poesio et al. (2004) as baseline. Since they did not explain whether they used the mention-mention or mention-entity model, we assume they treated antecedents as entities and use a 2 and 5 sentence window for candidates[8]. Since the GoogleAPI is not available any more, 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 as well as by replacing proper names. All other features (Table 2 in Section 5.1.1) are extracted as Poesio et al. did. A Naive Bayes classifier with standard settings in WEKA (Witten and Frank, 2005) is used. In order to evaluate their model in the more realistic setting of our experiment, we apply the *best first* strategy to select the antecedent for each anaphor.

6.3 Pairwise models

Pairwise model I: We use the *preposition pattern* feature ($feat1$) plus Poesio et al.'s salience features (Table 2). We use a 2 sentence window as it performed on a par with the 5 sentence window in the baseline. We replace Naive Bayes with SVM^{light} because it can deal better with imbalanced data[9].

Pairwise model II: Based on *Pairwise model I*. Local features $feat2$-$feat8$ from Table 2 are added.

Pairwise model III: Based on *Pairwise model II*. We apply a more advanced antecedent candidate selection strategy, which allows to include 77% of NP antecedents compared to 71% in *Pairwise model II*. For each anaphor, we add the top k salient entities measured through the length of the coreference chains (k is set to 10%) as additional antecedent candidates. For potential set anaphors (as automatically determined by pronoun or number heads), singular antecedent candidates are filtered out. We compiled a small set of adjectives (using FrameNet and thesauri) that indicate spatial or temporal sequences (see Example 3). For anaphors modified by such adjectives we consider only antecedent candidates that have the same semantic class as the anaphor.

[8]They use a 5 sentence window, because all antecedents in their corpus are within the previous 5 sentences.

[9]The SVM^{light} parameter is set according to the ratio between positive and negative instances in the training set.

6.4 MLN models

MLN model I: MLN system using local formulas $f1$-$f2$ and $f14$-$f20$. The same strategy as in *Pairwise model III* is used to select local antecedent candidates E_m for each anaphor m.

MLN model II: Based on *MLN model I*, all formulas in Table 1 are used.

6.5 Results

Table 4 shows the comparison of our models to baselines. Significance tests are conducted using McNemar's test on overall accuracy at the level of 1%.

		acc
improved baseline	*2 sent. + NB*	18.85
	5 sent. + NB	18.40
pairwise model	*pairwise model I*	29.11
	pairwise model II	33.94
	pairwise model III	36.35
MLN model	*MLN model I*	35.60
	MLN model II	**41.32**

Table 4: Results for MLN models compared to pairwise models and baselines.

MLN model II, which is inspired by the linguistic observation that globally salient entities are preferred as antecedents, performs significantly better than all other systems. The gains come from three aspects. First, by selecting the antecedent for each anaphor from the antecedent candidate pool E in the whole document 91% of NP antecedents are accessible compared to 77% in *pairwise model III*. Second, we leverage semantics and salience by using local formulas and discourse level formulas. Local formulas are used to capture semantic relations for bridging pairs as well as surface and syntactic constraints. Global formulas resolve several bridging anaphors together, often to a globally salient antecedent beyond the local window. Third, the model allows us to express specific relations among bridging anaphors and their antecedents ($f11$-$f13$).

However, our *pairwise model I* already outperforms *improved baselines* by about 10%, which suggests that our *preposition pattern* feature can capture more diverse semantic relations. The continuous improvements shown in *pairwise model II* and *pairwise model III* verify the contribution of our other features and advanced antecedent candidate selection strategy. *pairwise model III* would become too complex if we tried to integrate discourse level formulas $f7$, $f11$-$f13$ into antecedent candidate selection. *MLN model II* solves this task elegantly.

6.6 Discussion and error analysis

We analyse our best model (*MLN model II*) and compare it to the best local one (*pairwise model III*).

Anaphors with long distance antecedents are harder to resolve. Table 5 shows the comparison of correctly resolved anaphors with regard to anaphor-antecedent distance. We can see that the global model is equal or better to the local model for all anaphor types but that the difference is especially large for anaphora with antecedents that are 3 or more sentences away due to the use of global salience and accessibility of possible antecedents beyond a fixed window-size.

	# pairs	*MLN II*	*pairwise III*
sent. distance			
0	175	48.57	45.14
1	260	34.62	35
2	90	47.78	43.33
$\geq$3	158	**35.44**	16.46

Table 5: Comparison of the percentage of correctly resolved anaphors with regard to anaphor-antecedent distance. Significance tests are conducted using McNemar's test at the level of 1%.

We now distinguish between ”sibling anaphors” (anaphors that share an antecedent with other bridging anaphors) and ”non-siblings” (anaphors that do not share an antecedent with any other anaphor). The performance of our *MLN model II* is 54% on sibling anaphors but only 24% on non-sibling anaphors. This shows that our use of global salience and links between related anaphors does indeed help to capture the behaviour of sibling anaphors.

However, our global model is good at predicting the right antecedent for sibling anaphors where the antecedent is globally salient but not as good for sibling anaphors where the (shared) antecedent is a locally salient subtopic. Thus, in the future we need to model equivalent constraints for local salience of antecedents, taking into account topic segmentation/shifts to improve over the 54% for sibling anaphors.

The semantic knowledge we employ is still insufficient. Typical cases where we have problems are: (i) cases with very context-specific bridging relations. For example, in one text about the stealing of Sago Palms in California we found *the thieves* as a bridging anaphor with the antecedent *palms*, which is not a very usual 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 a later anaphor *the veto* could be the veto of either bills. Integration of the wider context apart from the two NPs is necessary in these cases. This includes the semantics of modification, whereas we currently consider only head noun knowledge. An example is that 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. These cases are not correctly resolved in our current model as we only extract NP phrases as potential candidate antecedents.

7 Conclusions

We provide the first reasonably sized and reliably annotated English corpus for bridging resolution. It covers a diverse set of relations between anaphor and antecedent as well as all anaphor/antecedent types. We developed novel semantic, syntactic and salience features based on linguistic intuition. Inspired by the observation that salient entities are preferred as antecedents, we implemented a global model for antecedent selection within the framework of Markov logic networks. We show that our global model significantly outperforms other local models and baselines. This work is – to our knowledge – the first bridging resolution algorithm that tackles the unrestricted phenomenon in a real setting.

Acknowledgements. Yufang Hou is funded by a PhD scholarship from the Research Training Group *Coherence in Language Processing* at Heidelberg University. Katja Markert receives a Fellowship for Experienced Researchers by the Alexander-von-Humboldt Foundation. We thank HITS gGmbH for hosting Katja Markert and funding the annotation. We thank our colleague Angela Fahrni for advice on using Markov logic networks.

References

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

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. LDC2006T13, Philadelphia, Penn.: Linguistic Data Consortium.

Aoife Cahill 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,* Seoul, Korea, 5–6 July 2012, pages 232–236.

Tommaso Caselli 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,* Genoa, Italy, 22–28 May 2006.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing,* Cambridge, Mass., June 1975, pages 169–174.

Pedro Domingos and Daniel Lowd. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.

Angela Fahrni and Michael Strube. 2012. Jointly disambiguating and clustering concepts and entities with Markov logic. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pages 815–832.

Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7:395–433.

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

Matthew Gerber and Joyce Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):756–798.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language,* 69:274–307.

Donna Harman and Mark Liberman. 1993. TIPSTER Complete. LDC93T3A, Philadelphia, Penn.: Linguistic Data Consortium.

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

Emmanuel Lassalle 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),* Faro, Algarve, Portugal, 6–7 October 2011, pages 35–46.

Katja Markert, 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*. Budapest, Hungary, 14 April 2003, pages 39–46.

Katja Markert, 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,* Jeju Island, Korea, 8–14 July 2012, pages 795–804.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 155–163.

Shachar Mirkin, 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*, Uppsala, Sweden, 11–16 July 2010, pages 1209–1219.

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

Malvina Nissim, Shipara Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pages 1023–1026.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. LDC2011T07.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

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

Massimo Poesio. 2003. Associate descriptions and salience: A preliminary investigation. In *Proceedings* *of the EACL Workshop on the Computational Treatment of Anaphora*. Budapest, Hungary, 14 April 2003, pages 31–38.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.

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

Altaf Rahman 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*, Avignon, France, 23–27 April 2012, pages 798–807.

Sebastian Riedel. 2008. Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, 9–12 July 2008, pages 468–475.

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

Josef Ruppenhofer, 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)*, Uppsala, Sweden, 15–16 July 2010, pages 45–50.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of STARSEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Montréal, Québec, Canada, 7–8 June 2012, pages 1–10.

Wee Meng Soon, 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.

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

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

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, Cal., 2nd edition.