Distributional Semantics for Resolving Bridging Mentions

Tim Feuerbach and **Martin Riedl** and **Chris Biemann**
Language Technology Group, CS Department, TU Darmstadt, Germany
uni@spell.work, {riedl, biem}@cs.tu-darmstadt.de

Abstract

We explore the impact of adding distributional knowledge to a state-of-the-art coreference resolution system. By integrating features based on word and context expansions from a distributional thesaurus (DT), automatically mined IS-A relationships and shallow syntactical clues into the Berkeley system (Durrett and Klein, 2013), we are able to increase its F1 score on bridging mentions, i.e. coreferent mentions with non-identical heads, by 8.29 points. Our semantic features improve over the Web-based features of Bansal and Klein (2012). Since bridging mentions are a hard but infrequent class of coreference, this leads to merely small improvements in the overall system.

1 Introduction

Automatically recognizing coreference – relating lexical items that refer to the same entity or context in a text – is an important semantic processing step for text understanding tasks such as fact extraction, information retrieval, and entity linking.

A common problem of coreference systems is their inability to resolve bridging mentions, i.e. coreferent mentions with non-identical heads (Vieira and Poesio, 2000). For example, a system requires semantic knowledge to detect the hypernymic relationship that holds between mentions like *a preliminary agreement* and *the pact*. Similarly, modeling selectional preference relies on information beyond the pronoun context itself.

There are two different kinds of approaches employed in the past to make this knowledge available as features to a coreference resolution system. The first class uses manually crafted resources like WordNet or Wikipedia (Poesio et al., 2004; Ponzetto and Strube, 2006). Despite their quality, they may decrease the performance when added to the system (Lee et al., 2011; Zhou et al., 2011). Further disadvantages are their limited size, slow growth and general-purpose nature. In contrast, using unsupervised/semi-supervised methods for generating knowledge is only limited by the size of input data and adapts to the target domain.

We present features exploiting automatically obtained distributional knowledge, following the distributional hypothesis formulated by Harris (1954) that words in similar contexts bear similar meanings. For that we resort to a distributional thesaurus (DT; Lin, 1998) listing semantically similar terms, as well as hyponym-hypernym relations (IS-As) acquired with Hearst patterns (Hearst, 1992), both made available by the JoBimText Project (Biemann and Riedl, 2013). When added to the state-of-the-art Berkeley Coreference Resolution System (Durrett and Klein, 2013), these features show a significant positive impact on bridging mentions.

2 Related Work

Our work is very similar to Bansal and Klein (2012), who created, among others, features based on IS-As, distributional clusters, and pronoun contexts. However, we chose to use a DT's list of similar words instead of clustering, and dependency relations as context features instead of N-gram neighborhood. We will compare our approach to Bansal and Klein's features below.

Distributional methods for coreference resolution are mostly pattern-based (Haghighi and Klein, 2009; Kobdani et al., 2011). Recent work by Recasens et al. (2013) used news events as context and exploited rewordings of the same story in different sources.

Semantic similarity for the resolution of bridging mentions has been employed by Poesio et al. (1998), Gasperin et al. (2004), and Versley (2007), yet all three works are applied to oracle anaphoric

mentions, thus not facing spurious mentions, i.e. phrases that are non-referring in the gold standard. Ng (2007) and Lee et al. (2012) made use of Lin's theasurus in a fully-featured system, but with a smaller expansion size (5 and 10 words, respectively).

3 Method

We added our features to the state-of-the-art Berkeley Coreference Resolution System (Durrett and Klein, 2013), which also acts as our baseline. It employs a mention-pair model by assigning each predicted mention a latent antecedent. The probability of a mention m having antecedent a is estimated using a log-linear model and competes with the likelihood of m being non-anaphoric. Features are binary and distinguished between features on mention pairs and features on anaphoricity resp. the candidate antecedent.

For our experiments, we used the system's FINAL feature set. Regarding anaphoricity and the candidate antecedent, it uses the mention's size in words, syntactic uni- and bigrams of the head, as well as lexicalizations of the head, first, last, preceding, and following word as features. Pairwise features are the distance between the two mentions, once as the number of sentences and once as the number of mentions; whether one mention is within the boundaries of the other; whether they belong to the same speaker; the candidate antecedent's number and gender using data by Bergsma and Lin (2006); the syntactic uni- and bigrams of both mentions; mention string match or containment; head string match or containment. See Durrett and Klein (2013) for a detailed description of the feature set.

The Berkeley System expands the feature space by feature conjunctions: If a pairwise feature f fires for current mention m_c and antecedent mention m_a, features $f \wedge \textit{type}(c)$ and $f \wedge \textit{type}(c) \wedge \textit{type}(a)$ are also activated, where type$(\cdot)$ returns a mention type literal based on the head's POS. For pronouns, this is the citation form; for proper and common nouns, PROPER and NOMINAL are returned, respectively.

Our distributional knowledge comes from a DT. Biemann and Riedl (2013) generalized Lin's thesaurus (Lin, 1998) by distinguishing between *terms* (e.g. words) and *context features* (e.g. dependency relations). The *holing operation* @ extracts terms and features from surface text and is used both for training and querying the DT. The DT lists for each term the n semantically most similar terms, where n is the expansion size parameter, and semantic similarity is defined as the number of shared significant contexts.

4 Experimental Setting

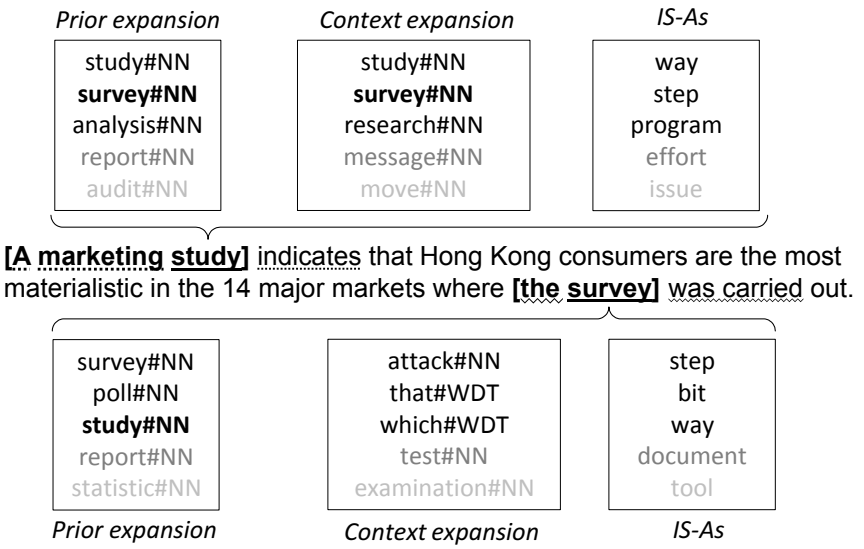
We adapted the training and evaluation data and splits from the CoNLL-2011 shared task on coreference resolution (Pradhan et al., 2011), which contains 2,999 documents from the OntoNotes v4.0 corpus (Hovy et al., 2006), and took the number and gender data from the task. Training and testing was performed with predicted mentions on the AUTO set of automatically preprocessed documents. We used the Berkeley System in version 1.0 and a DT created from 120M sentences of news texts ($n = 200$) using a dependency parse holing system (Biemann and Riedl, 2013, 72 f.) and including IS-As clustered into senses (Gliozzo et al., 2013).[1] Its terms are composed of a single word's lemma and its POS tag (e.g. *pact#NN*), while context features are neighbor terms in a dependency parse, complemented by the dependency label and governing direction (e.g. *governing#amod#preliminary#JJ*).

For evaluation, we used the standard coreference metrics MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (entity/ϕ_4-CEAF; Luo, 2005), as well as their average, computed with the reference scorer v7 (Pradhan et al., 2014).

Additionally, we evaluate precision and recall on *system-bridging mentions*,[2] i.e. mentions that appear as bridging to the system, but not necessarily to a human. Let *head*(m_i) return the predicted head of the i-th mention in a document, $C(m_i)$ be the (gold or system) coreference chain of m_i, and $C^*(m_i) = \langle m_j : m_j \in C(m_i) \wedge j < i \wedge \textit{head}(m_j)\ \textit{is a noun} \rangle$ be the sequence of noun antecedents of m_i. A mention m_i is *system-bridging* if head(m_i) is a noun, $C^*(m_i) \neq \emptyset$, and for all $m \in C^*(m_i)$ it holds that *head*$(m) \neq$ *head*(m_i). A bridging mention m_i from the gold chain C_G is a true positive (tp) if m_i and its immediate predecessor from $C^*_G(m_i)$ are members of the same system entity, and a false negative (fn) otherwise.

[1] model downloaded from `http://sourceforge.net/projects/jobimtext/files/data/models/en_news120M_stanford_lemma/`

[2] Our definition is based on *quasi-bridges* from the Berkeley System's source code (Durrett and Klein, 2013).



Feature values: PRIOR(t_1,t_2) = 2, PRIOR(t_2,t_1) = 3, SHARED_PRIOR = 0.4, IS-IS-A(t_1,t_2) = false, IS-IS-A(t_2, t_1) = false, SHARED_IS-As = 0.7, IN_C-EXPANSION(t_1, t_2) = 2, IN_C-EXPANSION(t_2, t_1) = 13.

Figure 1: Expansions and feature values for an example pair of bridging mentions from the development set. Dotted and wavy lines indicate dependency relations used in the context expansion.

A mention m_i' is considered a false positive (fp) if it is bridging in the system chain C_S, but is not coreferent with its immediate predecessor from $C_S^*(m_i')$ in the gold standard.

5 Additional Features

We added pairwise features from four different categories to the system, of which the last one (attribute features) is only loosely tied to a DT. Rank-based features have been discretized using equal-width binning (bin size: 20), though values from the interval $[-2, 20]$ were spelt out explicitly. Real values from the interval $[0, 1]$ were discretized by simply rounding to the first decimal digit. In the following feature description, t_1 and t_2 denote the heads of the current and antecedent candidate mention in term form. Each asymmetrical feature has an additional instance with t_1 and t_2 reversed. Furthermore, the function $\text{expansion}(\cdot)$ takes a term as its argument and returns the 200 most similar terms according to the DT. The position of a term t in an expansion is reported by $\text{rank}(t, \cdot)$.

1. **Prior** features target a head word's list of semantically similar terms as returned by the DT's expansion.

- PRIOR: Its value is 0 if $t_1 = t_2$, -2 if $\text{expansion}(t_2) = \emptyset$, -1 if $t_1 \notin \text{expansion}(t_2)$, and $\text{rank}(t_1, \text{expansion}(t_2))$ otherwise.
- SHARED_PRIOR: The overlap of two expansions: $(|\text{expansion}(t_1) \cap \text{expansion}(t_2)|) / \min(|\text{expansion}(t_1)|, |\text{expansion}(t_2)|)$.

2. **IS-A** features operate on open class head words' hypernyms. To keep things simple, we treated all clusters equal.

- IS-IS-A: *True* if t_1 is among any of the IS-As of any cluster of t_2, *false* otherwise.
- SHARED_IS-As: Calculates the Dice index (Dice, 1945) between each IS-A cluster of t_1 and each of t_2 and returns the maximum value.

Since the data contains some noisy IS-As like *bit* (originating from *is a bit*), we added an additional lexicalized feature for SHARED_IS-A = *true* with the shared IS-A that has the highest frequency in the model.

3. A feature targeting the **context** of a mention's head to model selectional preference. For this, we define a context-based expansion (C-expansion). Similar to verb argument expectations (Lenci, 2011), we compose a list of the most likely words appearing in a given context, but do not restrict ourselves to verbs. We exploit the fact that term-context pairs are provided in the JoBimText model (Biemann and Riedl, 2013). Let C be the set of context features of a mention head in the text

		MUC			B^3			$CEAF_e$			
		P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*Average*
Development	BASELINE	69.88	63.25	66.40	61.86	52.98	57.08	57.69	54.31	55.95	59.81
	DUMMY	69.81	63.92†	66.74†	61.86	53.74†	57.52†	58.03	55.23†	56.60†	60.28†
	P	69.62	63.98†	66.68†	61.85	53.91†	57.60†	58.11	55.35†	56.69†	60.33†
	PI	69.73	64.12†	66.81†	62.00	54.13†	57.80†	58.24†	55.42†	56.79†	60.47†
	PIC	69.60	64.13†	66.76†	62.00	54.14†	57.81†	57.99	55.56†	56.75†	60.44†
	PICA	69.59	**64.54**†	**66.97**†	61.90	**54.73**†	**58.09**†	**58.31**	**56.04**†	**57.15**†	**60.74**†
	B&K (2012) CO	**70.09**	63.48	66.62	**62.77**†	53.35	57.68†	58.22†	54.69	56.40†	60.23†
	—"— +DUMMY	69.78	64.19†	66.87†	62.23	54.08†	57.87†	58.02	55.21†	56.58†	60.44†
Test	BASELINE	69.69	65.98	67.79	58.68	53.59	56.02	54.31	53.88	54.09	59.30
	PICA	69.17	**66.87**†	**68.00**	57.77	**54.49**†	56.08	**54.45**	**54.44**†	**54.44**	**59.51**
	B&K (2012) CO	69.30	66.11	67.67	58.10	53.62	55.77	54.31	53.63	53.97	59.14
	—"— +DUMMY	68.57	66.56†	67.55	57.12	54.12†	55.58	53.70	53.80	53.75	58.96

Table 1: Metric results achieved by the baseline, dummy setting, and incrementally adding features to the baseline (P = prior expansion, I = IS-A, C = C-expansion and A = attribute features). Also comparing to the Bansal and Klein (2012) co-occurrence feature. Scores with a dagger (†) are significantly better than the BASELINE (paired bootstrap resampling test with $N = 10000$ and $p = 0.05$ (Koehn, 2004)).

and $T = \{t_1, \ldots, t_n\}$ the set of terms for which there exists a $c_j \in C$ such that the pair (t_i, c_j) is a member of the model. We sort the members of T in the descending order of their probability $P(t_i|C)$ and take the first 200 elements as the target term's C-expansion. Defining $P(t_i|C)$, we assume conditional independence and calculate the plus-one-smoothed MLE as $\prod_{c_j \in C}(sig(t_i, c_j) + 1)/(V + \sum sig(*, c_j))$, with $sig(\cdot, \cdot)$ returning the significance value of a term-feature pair stored in the model, and V as the vocabulary size. The coreference feature IN_C-EXPANSION then returns the rank of t_1 in t_2's C-expansion with PRIOR's result semantics. If t_1 is from a closed word class, it is first mapped to the first open word class term from its own C-expansion. Unlike typical takes on selectional preference, we expand all mention heads, not only pronouns, to take their *contextual role* (Bean and Riloff, 2004) into account and to have at least some semantic knowledge for out-of-vocabulary terms.

4. **Attribute** features inspired by Vieira and Poesio (2000, 556 f.;560) guessing properties of mentions from dependency relations in the text. We consider as attributes all words in a copula, appositive, relative clause, or compound relation to a mention's head and added the following features:

- ATTR_PRIOR = {*no attributes*, −2, …, 200}: Expands t_2, looks up each attribute of t_1 in t_2, and reports the best rank as in PRIOR.
- ATTR_IS-IS-A = {*true*,*false*}: Its value is *true* if t_1 is among any IS-A set of any attribute of t_2, *false* otherwise. If *true*, adds an additional version with the lexicalized IS-A.

Figure 1 illustrates the first three feature groups by means of a sentence from the development set. While the baseline treats those mentions as separate entities, our distributional features lead to their correct resolution.

6 Results

We present the results[3] of the modifications in Table 1. We also compare to a dummy system with the full feature set whose prior expansions return the identity, while the C-expansion and IS-A clusters are empty. This system profits from lemmatization as well as the syntactic clues provided by the attribute features.

While the BASELINE was unable to solve the introductory example, the distributional features provide the system enough confidence in assigning the mentions with the non-identical heads *pact* and *agreement* to the same entity. C-expansions had only low impact on performance. In a manual analysis, we observed many cases in which the sematically less preferable antecedent was selected, or in which non-coreferent pronouns were assigned to an entity, for example linking *you* in

[3] differences to reported scores in (Durrett and Klein, 2013) due to corrections of errors in the scoring script, see (Pradhan et al., 2014)

Thank you for your visit to a previous occurrence of *God* because of the common phrase *Thank God*. In comparison to PI, the recall on singleton pronouns decreased by 1 point, while the pairwise recall on anaphoric pronouns increased only by 0.4 points.

The final results on the test set in Table 1 were obtained by training on the conjunction of training and development data. We sacrifice some precision for better recall. Unfortunately, the increase in average F1 is not significant.

For comparison, we also integrated the feature set by Bansal and Klein (2012), computed on the Google Web N-gram corpus (Brants and Franz, 2006), into the Berkeley system. It includes the following features: *General co-occurrence* targets the general frequency of two head words appearing near to each other. *Hearst* works like our IS-IS-A feature. *Entity-based context* collects lists of seeds *y* in the pattern `h (is|are|was|were) (a|an|the)? y` in decreasing order of frequency, and reports whether there is a match in the top k seeds of the two head words. It also returns the dominant POS of the matched words. *Pronoun context* substitutes pronouns with their antecedent and estimates the likelihood of the new sequence. Finally, the *cluster* feature returns the sum of the earliest match positions of the two headwords' cluster ID lists, using phrasal clusters obtained by Lin et al. (2010).

We experimented with different permutations of these features, including the sets proposed in Bansal and Klein (2012), but found a set containing only the co-occurrence feature to perform best with regards to the average metrics score.[4] The results can be found in Table 1 noted as *B&K*. Remarkably, the feature rather increases precision than recall. The cluster feature led to a performance decrease already on the development set. This may stem from the many semantically unrelated word pairs, like *swords – elephants* or *definition – horror*, which share the same top cluster.

The models' results on bridging mentions are displayed in Table 2. We outperformed the baseline on both sets (F1 increased on test by 8.29 points). The positive impact on the metric scores is minor though, since only 7.6% of all mentions in the development set are bridging.

Again, we compare to the Bansal and Klein (2012) features, this time choosing the set performing best with regards to bridging mentions, which contains all features except *pronoun context*, which achieved an increase of 3.69 absolute F1 points on the test set. To assess whether these features are subsumed by our set or provide additional value, we also show the results of combining both in Table 2. The decrease in precision on the test set suggests that the Web features introduce too much noise to the system.

[4]For binning, we tried bin sizes 1, 0.5, and 0.25. For the entity features, we tried $k \in \{10, 20, 50, 100, 200\}$.

Bridging	**P**	**R**	**F$_1$**
BASELINE-Dev	36.21	15.51	21.72
DUMMY-Dev	41.36	17.45	24.55
PICA-Dev	**44.87**	23.82	31.12
B&K (2012)*-Dev	39.15	19.67	26.18
B&K (2012)*+Dummy-Dev	42.81	21.98	29.04
B&K (2012)*+PICA-Dev	44.19	**24.56**	**31.57**
BASELINE-Test	38.06	17.32	23.81
PICA-Test	39.47	27.05	**32.10**
B&K (2012)*-Test	37.97	21.56	27.50
B&K (2012)*+PICA-Test	36.84	**27.33**	31.38

Table 2: Precision, recall and F_1 scores on bridging mentions. Bolded improvements are significant over the baseline ($p = 0.05, N = 10000$).

7 Error Analysis

Error	BASELINE	PICA	Δ
Span	**399**	404	+5
Conflated entities	**1303**	1319	+16
Divided entities	1626	**1593**	-33
Extra entities	**521**	559	+38
Missing entities	881	**820**	-61
Extra mention	**577**	618	+41
Missing mention	862	**842**	-20

Table 3: Development set error counts comparison

As shown by an automatic classification of errors by the Berkeley Coreference Analyser (Kummerfeld and Klein, 2013) in Table 3, our system is prone to create spurious entities and mentions. The problem arises from semantic relations in the DT that are actually indicators of non-coreference (e.g. antonymy, co-hyponymy), but nevertheless ranked high. The similarity measure does not differentiate between these relations. This produced links like *Taipei – South Korea* and *the men – the women*. Since the hypothesis that a mention is non-referring has low probability if it begins with a determiner, the system desperately "searches" for an antecedent. Because of our semantic features, the system achieves higher confidence in

	ACR	ATT	CAN	DAT	DISC	HEAD	HYP	TATT	LEM	MET	SYN	INV	$\sum$
BASELINE	2	77	0	2	1	26	46	0	0	6	5	3	
PICA	3	99	1	3	4	31	97	1	1	6	9	3	
Total	23	235	50	32	171	58	409	13	10	38	32	12	1083

Table 4: Comparison of the numbers of resolved bridging mentions in the development set, broken down per type.

linking mentions with diverse heads if they bear at least some semantic similarity, creating spurious chains, which is punished by MUC and B^3 precision.

This intuition is backed by a manual analysis we conducted on 100 random errors not made by the baseline. When examining each of the system's antecedent decisions and their weights, we found that 23% of the wrong links were chosen because of distributional semantics features. The majority of these semantic errors were triggered by the PRIOR feature, whereas only one of them could be ascribed to the IS-A feature. Here, the recall-oriented clustering of IS-As in the DT (Gliozzo et al., 2013) produced an incorrect hypernymic relation between *Chaidamun Basin* and *the country*.

We classified the 1083 bridging mentions from the development set according to the knowledge required for resolution or their semantic relationship with a previous mention into the following categories:

ACR One head is an acronym of the other.

ATT One head is an attribute of the other as defined in Section 5.

CAN One head is in a CAN-BE relationship with the other, e.g. *pilot* and *man*.

DAT Temporal deixis like *today – the 30th*. We attribute the low recall of this class to the fact that the Berkeley system's FINAL feature set does not make use of named entity labels.

DISC Bridging mentions requiring textual entailment techniques, e.g. *my mother* and *Thelma Wahl*, sophisticated world knowledge as in the case of *Martha Stewart – the comeback queen*, or a discourse model to identify speakers or deixis.

HEAD Both heads are identical, but the system's head detection made a mistake. A large portion of these cases were Asian names, where the family name precedes the first name, and thus a strategy selecting the last word falls flat.

HYP The head of one mention is a hyponym of the other.

TATT Transitive attributes, i.e. one head is a hypernym, hyponym, acronym or synonym of one of the other mention's attributes, e.g. *Doctor Hunter* and *the physician*.

LEM Both heads have the same lemma.

MET The heads are in a metonymous relationship, e.g. *the Japanese government* and *Japan*.

SYN The heads are synonyms or near-synonyms, e.g. *dad* and *father*. This class also contains spelling variants and typos of proper names.

INV Invalid: At least one mention's head is a pronoun, but does not have the appropriate POS tag.

The results of both systems are shown in Table 4. Except for INV and MET, we increased the number of recalled mentions across all types. Hypernymic relationships form the largest class, making up more than a third of all system bridges in the development set. This was also the category with the strongest improvement: the number of recalled mentions doubled from 46 to 97 (23.7% of class size). We found that IS-A features are not solely responsible for this increase. For example, IS_IS-A did not fire for the links *a marketing study – the survey*, *the balloting – the elections* and *the insurrection – the Oct. 3 failed coup in Panama*, which were resolved thanks to the prior expansion. On the other hand, bridging mentions with attributes in transitive relationships, which inspired our attribute features, form only a small class with 13 members, from which we resolved 1 (baseline: 0).

8 Conclusion and Outlook

We have shown that our DT-based approach adds more than double the amount of absolute F1 points on bridging mentions in the test set than the semantic features described by Bansal and Klein (2012). However, undesired semantic relations present in the DT lead to a decrease in general resolution precision. A possible solution are asymmetrical *directional similarity measures* (Lenci, 2014) which bring preferred semantical relations to the top of the expansion, thus allowing the system to assign higher weights to these ranks. Also, classifiers using entity-mention or ranking models may profit from directly comparing ranks instead of learning separate weights like in the case of the Berkeley system's mention-pair model. While our results confirm that introducing semantic features in a coreference system is an "uphill battle" (Durrett and Klein, 2013), we have shown positive impact on a hard class of coreference using automatically acquired semantic information instead of manually constructed lexical resources. This will enable more domain-adaptive coreference resolution systems in the future, as well as open up avenues for adding semantic features for low-resourced languages.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proc. Linguistic Coreference Workshop at LREC*, pages 563–566, Granada, Spain.

Mohit Bansal and Dan Klein. 2012. Coreference semantics from web features. In *Proc. ACL*, pages 389–398, Jeju Island, Korea.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proc. NAACL*, pages 297–304.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proc. ACL*, pages 33–40, Sydney, Australia.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Thorsten Brants and Alex Franz. 2006. The Google Web 1T 5-gram corpus version 1.1. *LDC2006T13*.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proc. EMNLP*, pages 1971–1982, Seattle, WA, USA.

Caroline Gasperin, Susanne Salmon-Alt, and Renata Vieira. 2004. How useful are similarity word lists for indirect anaphora resolution? In *Proc. DAARC*, S. Miguel, Azores, Portugal.

Alfio Gliozzo, Chris Biemann, Martin Riedl, Bonaventura Coppola, Michael R. Glass, and Matthew Hatem. 2013. JoBimText Visualizer: A graph-based approach to contextualizing distributional similarity. In *Proc. 8th TextGraphs at EMNLP*, Seattle, WA, USA.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proc. EMNLP*, pages 1152–1161, Singapore.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. COLING*, volume 2, pages 539–545, Nantes, France.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proc. NAACL-HLT*, pages 57–60, New York, NY, USA.

Hamidreza Kobdani, Hinrich Schütze, Michael Schiehlen, and Hans Kamp. 2011. Bootstrapping coreference resolution using word associations. In *Proc. ACL*, volume 1, pages 783–792, Portland, OR, USA.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395, Barcelona, Spain.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proc. EMNLP*, Seattle, WA, USA.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proc. CoNLL: Shared Task*, pages 28–34, Portland, OR, USA.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proc. EMNLP-CoNLL*, pages 489–500, Jeju, Korea.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proc. CMCL*, pages 58–66, Portland, OR, USA.

Alessandro Lenci. 2014. Will distributional semantics ever become semantic? Talk at the 7th International Global WordNet Conference, Tartu, Estonia. http://gwc2014.ut.ee/lenci_distributional_semantics_gwc2014.pdf.

Dekang Lin, Kenneth Ward Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, et al. 2010. New tools for web-scale n-grams. In *Proc. LREC*, Valletta, Malta.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. COLING*, volume 2, pages 768–774, Montreal, QC, Canada.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. HLT-EMNLP*, pages 25–32, Vancouver, BC, Canada.

Vincent Ng. 2007. Shallow semantics for coreference resolution. In *Proc. IJCAI*, pages 1689–1694, Hyderabad, India.

Massimo Poesio, Sabine Schulte im Walde, and Chris Brew. 1998. Lexical clustering and definite description interpretation. In *Proc. AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA, USA.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proc. ACL*, Barcelona, Spain.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. NAACL-HLT*, pages 192–199, New York, NY, USA.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proc. CoNLL*, pages 1–27, Portland, OR, USA.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proc. ACL*, pages 22–27, Baltimore, MD, USA.

Marta Recasens, Matthew Can, and Daniel Jurafsky. 2013. Same referent, different words: Unsupervised mining of opaque coreferent mentions. In *Proc. HLT-NAACL*, pages 897–906, Atlanta, GA, USA.

Yannick Versley. 2007. Antecedent selection techniques for high-recall coreference resolution. In *Proc. EMNLP-CoNLL*, pages 496–505, Prague, Czech Republic.

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

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. MUC-6*, pages 45–52, Columbia, MD, USA.

Huiwei Zhou, Yao Li, Degen Huang, Yan Zhang, Chunlong Wu, and Yuansheng Yang. 2011. Combining syntactic and semantic features by SVM for unrestricted coreference resolution. In *Proc. CoNLL: Shared Task*, pages 66–70, Portland, OR, USA.