Chapter 14
Using Lexical and Encyclopedic Knowledge

Yannick Versley, Massimo Poesio, and Simone Ponzetto

Abstract Semantic information is one of the indispensable ingredients that are necessary to raise the performance of anaphora resolution both for pronominal anaphors and for anaphoric definite descriptions beyond the baseline level. In contrast to hard criteria such as binding and agreement constraints, however, the question of semantic constraints and preferences and its operationalization in a system that performs anaphora resolution, is more complex and a larger variety of solutions can be found in practice.

14.1 Introduction

It is possible to construct a relatively well-performing coreference system that is purely based on the preprocessing results (syntax, named entity resolution) together with precise but ultimately shallow heuristics, e.g. the system of Lee et al. [39]. In this sense, distance and syntactic heuristics together with good animate/inanimate distinctions (see 12) for pronouns and string matching and string similarity heuristics give a system that performs quite respectably compared to simpler machine learning approaches such as that of Soon et al. [75]. Indeed, at the CoNLL-2011 shared task, Lee et al.'s Stanford Sieve system performed better than systems with more impressive inference and feature approaches.

This fact is only seemingly at odds with the often-stated claim that successful coreference resolution has to depend on world knowledge: Indeed, the most successful CoNLL-2012 entry, Fernandes et al. [18] and, to give a much earlier exam-

Yannick Versley
University of Heidelberg, e-mail: versley@cl.uni-heidelberg.de

Massimo Poesio
University of Essex, e-mail: poesio@essex.ac.uk

Simone Ponzetto
University of Mannheim, e-mail: simone@informatik.uni-mannheim.de

ple, including features targeting more elaborate phenomena, Kameyama et al. [31], which performed best at the original MUC-6 competition, all successfully use more elaborate features. Simultaneously, though, we find work that finds absolutely no benefit to more elaborate features, such as Ng and Cardie [54], who explore a large number of features and achieved substantial gains over [75], but found lexical features based on WordNet to be non-helpful, or Kehler et al. [32], who claim that a well-performing coreference resolution (at least in their case) does not benefit from selectional preference information. Such work, in turn, coexisted with even earlier work such as Carter's SPAR [11] that emphasizes the importance of common-sense knowledge, or work such as Harabagiu et al. [26], who find large gains from doing extensive modeling of semantic relatedness using an extended version of WordNet.

One factor of this is the issue of evaluation, particular what we will call **non-realistic settings** that are less sensitive to low-precision resolution behaviour where actual usage in a component for coreference resolution would create (too) many false positives, whereas in realistic evaluation settings (as they are standard in most work done today), more cautious techniques allow more modest (but practically relevant) performance gains.

A second factor is that reductions of coreference to a classification problem, as found in early machine learning approaches such as [75] or [54] have to approximate the structured prediction task of finding coreference chains through binary decisions, making the addition of additional features to a system a non-trivial undertaking, whereas the most well-performing machine learning systems in use today [7, 16, 18] use approaches that are more closely modeling the actual resolution process, and by consequent can use a much larger and richer feature set than the older approaches.

A third factor is the development of **larger and better corpora** in the last twenty years and of **lexical and encyclopedic resources** in the last ten years, which are all instrumental in the resolution for less trivial links with a precision that is high enough to benefit realistic coreference resolution.

Let us therefore first gain a clearer picture of the specific phenomena in anaphora and coreference resolution that can benefit from lexical and encyclopedic knowledge, and discuss the resources available for this task, before discussing work that integrates this kind of knowledge in more detail.

14.1.1 Phenomena requiring Lexical and Encyclopedic Knowledge

In coreference resolution, lexical and encyclopedic knowledge could conceivably help the resolution of pronominal anaphora (*Clinton–she*), the resolution of nominals to either names (*Beijing–the city*) or other nominals (*the capital–the city*), as well as the resolution of names variations that are not detectable through string matching (*IBM–Big Blue*).

Pronoun anaphora

Much early work on pronoun resolution that is still well-known today works completely using agreement constraints and factors such as recency and syntactic salience that do not need semantic information. Similarly, the single most effective heuristic for resolving non-pronouns, string matching, can work completely without any semantic information. This prevalence of knowledge-poor techniques, however, is not due to accident, or ignorance of the problem: Early works on reference resolution such as Charniak's 1972 PhD thesis [12] or later Carter's SPAR system [11] explicitly acknowledge the importance of commonsense knowledge to the interpretation of pronouns and non-pronouns alike, to the point of mentioning examples like the following (due to Charniak [12]) where correct resolution of the pronoun requires a full understanding of the text:

(20) Today was Jacks birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a top.
Dont do that said Penny. Jack has a top. He will make you take it back.

In the example, it does not refer to the most recent top mentioned (the one Jack has), but the (hypothetical!) top that Janet wants to buy. While the full understanding of a text, which would be necessary to resolve such cases successfully, is still not within reach, verifying the fit of the antecedent with the context of the anaphor (e.g., *wear–shirt* being more likely than *wear–store*) or finding plausible progressions of contextual roles (when *A steals* something, and *B investigates*, *A* is more likely to be the one who *is arrested*, while *B* may be the one who *arrests*) seem to be in the range of an approach that uses sufficient lexical knowledge in an effective way.

To operationalize semantic constraints, we have at our disposition the anaphor and antecedent itself, but also their context (mostly, in the sense of the immediate syntactic context rather than a larger notion of discourse context, which would be much harder to model). For pronouns, the form of the anaphor itself (i.e., the pronoun) gives preciously few information beyond compatibility considerations. Therefore, most of the work on using semantic or world knowledge information in pronoun resolution has focused on using context elements, especially **selectional restrictions**. More ambitious approaches in this respect also try to exploit regularities in **event participants** in a discourse, for example the fact that an object of kidnap would occur later as an object (rather than the subject) of release, as in the following example (due to Bean and Riloff [3]):

(21) Jose Maria Martinez, Roberto Lisandy, and Dino Rossy, who were staying at a Tecun Uman hotel, were kidnapped by armed men who took them to an unknown place. After they were released...

However, such semantic preferences usually show **interaction with syntactic preferences**; for example, the following example is easy to misunderstand because a strong syntactic preference (unavailability of non-NP referents) works against the one plausible interpretation:

(22) After Windows 7 comes out in October, will Microsoft somehow force us XP users to stop using it?

The correct antecedent *XP* is dispreferred over the syntactically salient candidate *Windows 7*, requiring the reader to use world knowledge to infer that *it* in this case refers to Windows XP.

Definite Nominals

In the case of definite descriptions (common noun-headed noun phrases with a definite article, also called nominals), the information from the head of the anaphoric noun phrase is more useful and can in the ideal case be as good a filter for possible antecedents as shallow string matching can be, and most of the approaches in section 1.2 start from this idea. In the following example (cf. Versley, 2006 [81]), for example, we can exploit **lexical relations** that are good indicators of compatibility in an antecedent, as (female) pedestrian is a hyponym of woman, and car is a synonym of automobile:[1]

(23) An 88-year-old (female) pedestrian has been gravely injured in a collision with a car. When crossing the Waller Heerstrasse, the woman had obviously overlooked the automobile. The driver could not brake in time.

In a similar vein, we often find nominal anaphora where a name antecedent describes an instance of the concept (as in *Berlin ... the city*):[2]

(24) Even though Berlin ranks last when it comes to growth, the Senator for Economic Development already sees the city as a "Mekka for founders".

This neat connection between lexical relations and antecedence, however, does not mean that our problem is solved: on one hand, around half of the anaphoric definite descriptions do not have an antecedent with a clear[3] lexical relation.

On the other hand, a definite description can also designate a newly introduced entity which is either inferred from the general scenario or that has a non-identity anchor (see Chapter 13).

A different set of lexical relations such as those betwee *car* and *driver* can also enable **non-coreferent (associative) bridging relations**, which link an anaphoric definite description to a non-coreferent antecedent.

In an investigation of both coreferent and non-coreferent bridging relations between definite descriptions and mentions in the previous text, Poesio and Vieira [56] break up the bridging descriptions into six classes, motivated mostly by processing considerations:

---

[1] TBa-D/Z corpus, sentence 190

[2] translated from TBa-D/Z corpus, sentence 2015

[3] Clear is meant in the sense that it holds up to lexicographic criteria, as opposed to the contents of both terms being incomparable logically.

- lexical relations between the heads: synonymy, hypernymy, meronymy (e.g.: *new album ... the record*)
- instance relations linking definite descriptions to proper names (e.g.: *Bach ... the composer*)
- modifiers in compound nouns (e.g.: *discount packages ... the discounts*)
- event entities introduced by VPs (Kadane Oil Co. is currently *drilling ... the activity*)
- associative bridging on a discourse topic (*the industry* in a text on oil companies)
- more complex inferential relations, including causal relations

In the 204 bridging definite descriptions from Poesio and Vieira's corpus[4], 19% of definite descriptions had a lexical relation between common noun heads and 24% were definite descriptions referring to named entities.

Many of these *bridging descriptions* thus pertain to cases that are anaphoric, but which we do not want a system to annotate as coreferent. For example, given a house, we can talk about the door without otherwise introducing it as a referent), which is acceptable even when a semantically similar antecedent would be available.

In addition to the difference between coreference, semantic compatibility, lexical relations and relatedness/association, we have to keep in mind that different corpora have annotation guidelines that, although well-motivated, may strike the cursory reader as counterintuitive: The TBa-D/Z guidelines, for example, preclude **generic mentions** from annotation since their reference properties are not always clear-cut. Consider the following example 25:

(25) The pelage of the Siberian tiger is moderately thick, coarse and sparse [...] Generally, the coat of western populations was brighter and more uniform than that of the Far Eastern populations. [...] In the southeast Trans-Caucasus, the Siberian tigers main prey was wild boar.

In this example,[5] the Siberian tiger refers generically to the whole subspecies and various subpopulations, which would make annotation potentially difficult.

To provide a different example, [6] consider metonymic mentions of a country such as Israel for either the country, its government, or its population, such as in the following example

(26) Israel will ask the United States to delay a military strike against Iraq until the Jewish state is fully prepared for a possible Iraqi attack with nonconventional weapons, the defense minister said in remarks published Friday. [...] Israel is equipping its residents with gas masks and preparing kits with antidotes. [...] Israels armed forces chief, Lt. Gen. Amnon Lipkin-Shahak, appeared on national television Friday night in attempt to reassure a jittery public. When events are concrete and actually unfolding, we will

[4] a small subset of the Wall Street Journal section of the Penn Treebank

[5] from http://en.wikipedia.org/wiki/Siberian_tiger

[6] ACE-2, document NWIRE/APW19980213.1305

> include the entire population in the measures we decide are appropriate to take, Shahak said in an interview with Channel Two Television.

In the example, Israel and the entire population are coreferent, as they pertain to different aspects (government, population) of the same named entity; as in the example, this sometimes leads to undesirable results, as Israel, its, and its residents violates several strong linguistic constraints (i-within-i, c-command).

In the OntoNotes corpus, a different treatment of metonymy is used, and the Jewish state and its residents would always be separate entities, with counterintuitive effects when considering metonymic mentions such as in Israel is in fear, which would be non-coreferent to the mention in Israel will ask the United States, and the underspecified Israels armed forces would pose a problem in these guidelines.

Therefore, different kinds of semantic information would be needed for the ACE and OntoNotes corpora: For the former, one would need to find metonymous mentions such as *China–Beijing*, whereas for the latter, disambiguation would be needed to separate different uses of the same name *Israel*. Due to the fact that only a minority of definite descriptions is anaphoric, a successful resolver would integrate anaphoricity detection (see chapter 13) and the resolution proper. A corollary of this is that results for methods that improve the recall of coreference resolution through means of noisy features (for example, through unsupervised learning) often have to be taken with a large grain of salt, since many evaluation settings that are common in this area either presupposing a perfect filter for discourse-new mentions or only considering antecedent candidates that are themselves part of a coreference chain overestimate the utility of such features.

14.1.2 Lexical and Encyclopedic Information Sources

While earlier work such as Kameyama's MUC-6 system [31] relied on resources specifically built or compiled for the coreference system, and some resources such as gender information are specific to anaphora resolution (see Bergsma chapter), state-of-the-art systems that target lexical and encyclopedic knowledge heavily rely on general-purpose resources to provide lexical and encyclopedic information. Because this is partly independent of the approach for integrating the information into an anaphora or coreference system, we will briefly review the main contenders here.

Lexical resources

One of the oldest resources for lexical knowledge is WordNet [49], born out of a desire to be able to find words according to semantic criteria, which has become one of the staples in English-language natural language processing. The backbone of WordNet consists of *synsets* which link synonymous word senses together, and

which are in turn organized into a (mostly) taxonomic structure with a smaller number of non-taxonomic relations.

Using WordNet's structure, it is possible to find synonyms (*the suit*...*the lawsuit*) and hyperonymic relations (*the villa*...*the house*), but also some near-synonyms when looking at coordinate terms that have a direct hyperonym in common. It is possible to use these relations in a direct manner in order to detect antecedents for nominal mentions that have one of these well-known semantic relations, e.g. in Vieira and Poesio's [84] approach for resolving definite descriptions.

WordNet has been used in coreference resolution by defining semantic classes that encompass certain subtrees of the concept tree, as used in the system of Soon et al. [75], but it is also possible to use distance or similarity measures for coreference, as demonstrated by Lin's MUC-6 system [40], or in the experiments of Ponzetto and Strube [60].

In addition to taxonomic information, WordNet (and many non-English wordnets) offers non-taxonomic relations, which are often too sparse to be a reliable source of information, and glosses, which provide a natural-language explanation of the meaning of words and can be exploited for (low-precision) relatedness measures such as those used by Harabagiu et al. [26].

Another source of lexical (rather than encyclopedic) information can be found in the various resources that cover verbs in particular **FrameNet** [1] and **VerbNet** [33], which both help in generalizing over the grammatical role of one particular verb; because selectional preferences or co-occurrence of argument positions cannot be read off directly from FrameNet's or VerbNet's lexicon, they are most often used together to generalize verb-specific data rather than being used in isolation.

Semi-Structured and Structured Encyclopedic Knowledge

Purely lexical resources such as WordNet purposely do not cover named entities, and thus have long excluded all encyclopedic knowledge. The goal of providing a complete sense repository for English common nouns has led to disregard of information about individuals (i.e., named entities), as is evidenced by the limited number of individuals it contains only 9.4% according to Miller and Hristea [50] and by the fact that an individual-specific relation such as instantiation has been introduced with version 2.1 (i.e., in 2005).

As a consequence, earlier works such as Ng and Cardie [54] or Poesio et al.'s [58] investigation into this topic could only rely on the preprocessing results in addition to features extracted from raw text. However, this situation is changing with encyclopedic knowledge that is either extracted automatically or semi-automatically from Wikipedia, or (in the case of FreeBase) curated manually.

Wikipedia is a collaborative open source encyclopedia edited by volunteers and provides a very large domain-independent encyclopedic repository: the English version, as of December 2014, contains more than 4,675,000 articles with tens of millions of internal hyperlinks.

There are at least three main features which make Wikipedia attractive as a knowledge repository for AI and CL applications:

1. good **coverage** across many domains: it contains a large amount of information, in particular at the instance level
2. **multilingual**: it is available with a (mostly) uniform structure for hundreds of languages, even though the size of Wikipedias in different languages vary substantially: German, the second-largest edition of Wikipedia, still contains 1,789,000 articles and there are about 15 languages with more than 800,000 articles, and other languages such as Korean and Arabic, still have a sizeable number of articles (both around 300,000).
3. **up-to-date**: it includes continuously updated content, which provides current information.

Wikipedia exists only since 2001 and has been considered a reliable source of information for an even shorter amount of time [24], so researchers in CL have only later begun to exploit its content or use it as a resource. Since May 2004, Wikipedia contains a thematic categorization scheme by means of its categories: articles can be assigned to one or more categories, which are further categorized to provide a so-called category tree. In practice, this tree is not designed as a strict hierarchy, but contains a coexistence of multiple categorization schemes. Ponzetto and Strube [61], for example, posit that the category of hierarchies, while not being a taxonomy, is generally organized according to specificity.

Wikipedia also contains **structured information** beyond the categories: in particular, the attributes of many salient entities (plants, cities, US presidents, etc.) are listed in a standardized fashion in so-called *Infoboxes*, and there are also many tables with useful data (e.g. a table mapping countries to capitals, or demonyms such as Chinese or German to the respective country names), together with list pages that contain all entries of one particular category (e.g., *17th century composers*).

Thus, subsequent efforts to create resources such as **DBpedia** [48] worked on not just reshaping the category tree, but in extracting useful ontological information out of Wikipedia's data. Another undertaking, the **YAGO** knowledge base [78], sews together Wikipedia information with WordNet and the GeoNames gazetteer for place names. While not always meeting gold-quality standards (YAGO is based on an automatic extraction process, not manual annotation), YAGO yields a good combination of Wikipedia's good coverage for named entities and WordNet's taxonomical information.

YAGO includes about 80 relations (included LOCATEDIN for topological inclusion, or BORNIN for persons being born in a particular city), including a MEANS relation that relates names to their (potential) referent – the string "*Einstein*", for example, has MEANS relations to the concept nodes for Albert Einstein and the musicologist Alfred Einstein – and a TYPE relation which links a particular entitiy to classes they belong to (e.g., physicist). Through MEANS and TYPE relations, it is thus possible to get from a surface string to the corresponding entity concept as well as to attributes, or types, describing that entity.

The **FreeBase** database [8], which is partly based on data from Wikipedia but has since been enriched with data on many other entities in different domains; FreeBase is manually curated, and contains various properties and relations for each entity that are specific to each type of entity (i.e., persons have different properties than medical plants or rock bands).

Learning from Unannotated Text

Approaches using manually built knowledge bases rely on high-quality knowledge manually encoded by human experts at the cost of a (necessarily) limited coverage. By consequent, unsupervised (or even semi-supervised) techniques to learn relevant information from unannotated text – which, at least for general-domain text, exists in large quantities even in language other than English – are potentially very attractive since they would allow it to cover even less-frequent words. However, the most popular approaches in distributional semantics all suffer from relatively low precision, which makes it necessary to consider techniques that offer more precision (and, conversely, may be more modest in the coverage and achievable recall that they allow to achieve).

The most well-known among the techniques offering higher precision than purely considering word co-occurrences in large corpora consists in the extraction (and possibly weighting) of **lexicosyntactic patterns** within large corpora.

The occurrence of such patterns is taken to be indicative of particular lexical relations, for example the patterns introduced by Hearst [27] for hypernymy (e.g., *Y*s such as *X*, *X* and other *Y*s) or by Berland and Charniak [5] for part-of relations (*Y*s *X*, *X* of *Y*). These semantic relations can be then used to help identify strongly related mention pairs as coreferent.

Fleischman et al. [19] follow a similar approach, but extract a large number appositional patterns covering the following constructions based on POS patterns:

- Nominal/Noun constructions
  [*trainer*] [*Victor Valle*]
- syntactic appositions
  [*George McPeck*], [*an engineer from Peru*]

These patterns would individually have relatively low precision, which is why Fleischman et al. use the extracted patterns with a learning-based filter to improve the precision of the approach. To build the classifier-based filter, they use 5000 annotated pairs, together with features describing the surface form of nominal and noun, which allows them to filter the extracted apposition patterns using the learned classifier.

The **sources for unannotated text** that researchers have used for these purposes have varied with time: When, in the early 2000s, corpora such as the BNC counted as large, we soon saw both the coming of larger corpora – in particular, the *English Gigaword* corpus, which contains several billion words of newswire text, as well as Web corpora growing in size from one billion to several billions. Further, exploit-

ing the Web through search engine queries was identified as a useful technique by, e.g., Markert and Nissim [44], but is rarely used in modern systems because of the difficulties in scaling to larger corpora, as well as changing (or simply disappearing) APIs of search engines, which hinder both reproducibility and scalability of the approach. In contrast, some approaches that allow larger scale than normal Web corpora use the N-gram counts dataset published by Google[7], or very large Web corpora extracted from the ClueWeb dataset[8], which contains one complete Web crawl from 2009 and thus allows to construct large corpora in a more reproducible fashion.

Beyond using patterns found in a corpus, there is also the possibility to use standard approaches for **distributional similarity** or thematic relatedness, as used in some of the approaches in Section 14.2.1; However, *similarity* rather than *instance* relations between words tend to be less useful for at least the two following reasons:

- Most distributional "similarity" measures that are commonly used are actually *relatedness* measures that score highly on non-taxonomic pairs of words such as *house–door* or *currency–government*, which means that these measures will typically produce a large number of semantically dissimilar but thematically related spurious antecedents.
- In many cases, semantic similarity between noun phrases (such as between *US Software* and *Software from India*, or between a *goat* and a *sheep*) can still hold when the two mentions are incompatible, despite the fact that their context distributions are as similar as those of quasi-synonyms such as *home* and *house*.

Learning Semantic Information from Coreference Corpora

The fact that it is possible to learn semantics constraints or preferences for coreference from coreferentially annotated corpora sounds self-evident enough that one may wonder why only fairly recent systems do this. Part of the answer lies in the available corpora with coreference annotation (cf. Chapter 4): early coreference corpora were rather small (the MUC-6 and MUC-7 documents have below 100 documents together), which means that any such approach would run into data sparsity issues.

The second point that makes this an issue is the interaction between feature design and learning algorithm: the decision tree classifier used by, e.g. Soon et al. [75] or Ng and Cardie [54] works best when used with relatively few, informative, features, and the reduction of coreference to binary classification that was the dominant approach until recently does not necessarily benefit from large, sparse feature sets. Even classification approaches that work in a suitable machine learning framework and on large enough corpora, such as Daumé and Marcu's [15], or the ranking-based system for nominal and names of Versley [81], which could (or can) potentially ben-

[7] Brants, Thorsten, and Alex Franz. Web 1T 5-gram Version 1 LDC2006T13. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

[8] `http://lemurproject.org/clueweb09.php/`

efit from this information [insert graphics from thesis presentation] were designed with data scarcity in mind.

Modern corpora such as OntoNotes [62], TBa-D/Z [79] or the Prague Dependency Treebank [52] offer as much as (or more than) one million words of coreferentially annotated text, which means that certain facts (such as that *the country* likes named entity mentions of the GPE class as antecedents) can be learned effectively from the training set (which in turn may limit the applicability to other domains; the issue of out-of-domain performance of coreference systems has not been investigated as systematically as, e.g., the out-of-domain performance of statistical parsers).

14.2 Early Approaches

Much of the research until early 2000s targeted specific linguistic phenomena, for the simple reason that the available annotated corpora lent themselves more to manual inspection and careful modeling than to the construction of machine learning-based approaches that rely purely on the annotated data. Subsequent research presents results for all kinds of mentions, but often suffers from the fact that evaluation algorithms were described but not always implemented in the same way, or applied on the same kind of output (see the discussion in Chapter 5). In particular, many researchers used the gold-standard mentions as system input, which leads to a severe overestimation of the system's precision, and has been widely criticized both informally and in the literature [43, 76]; as this kind of evaluation has only a weak relation to system performance in practice, quantitative results from such comparison have to be taken with a large grain of salt; however, we find the explorative part of the work may be interesting in its own right.

14.2.1 Approaches for specific phenomena

Using Semantic Compatibility Information for Pronoun Resolution

One basic distinction that is necessary for resolving English pronoun anaphora is the male/female/inanimate distinction that corresponds to the *he*/*she*/*it* form of pronouns (and corresponds to a difference that is realized in the morphological properties of the antecedent noun phrase in morphologically richer languages). We refer the reader to Chapter 12 for a more extensive discussion of this, noting in passing that work such as Ge et al.'s generative model for pronoun anaphora [23] or Soon et al.'s distinction of semantic classes [75] have the important benefit of giving the system information about the he/she/it distinction.

The work of Dagan and Itai (1990) similarly presents an approach where automatic parses from a 60 million word corpus are used to extract statistics about

subject-verb and object-verb cooccurrences, which are then used as a model of **selectional preferences**. Using a hand-selected sample of it pronouns where the antecedent as well as one or more other candidates compatible in number, gender and syntactic position were in the same sentence, Dagan and Itai found that in 64% of the cases, antecedent and candidates all occurred at least five times in the parsed corpus, and of these, 87% had the correct antecedent allowed by their selectional preference model, and in about half of these cases, the antecedent was the only one that fits the selectional preferences.

While this approach clearly steers free of most problems that would hinder the use in a full coreference system among others, noise in the determination of agreement features, classification of named entities, or treatment of infrequent words it has certainly inspired further research that aims at using selectional preferences. For instance, Dagan et al. [13] present a post-processor for a rule-based pronoun resolver, which breaks ties in the systems coreference decisions based on predicate-argument cooccurrence statistics, i.e. how many times a pronoun occurs as the argument of a certain predicate. A model based on distributional methods is also presented in Klebanov and Wiemer-Hastings [34], which use Latent Semantic Analysis [37] to model world knowledge for pronoun resolution. Finally, Kehler et al. [32] discuss the integration of selectional preference features in a maximum-entropy based pronoun resolver; they find that in the absence of number or gender agreement features, selectional preference features give a very visible loss in accuracy, whereas otherwise they yield a small (but not statistically significant) improvement over a model with no selectional preferences.

Using Semantic Compatibility Information for Nominal Resolution

In the resolution of nominal coreference, Vieira and Poesio's [84] on a subset of the Penn Treebank texts and subsequent work using the GNOME corpus of definite descriptions [58] is focused on the question of resolving definite descriptions (i.e., nominal mentions with a definite article). Among the anaphoric definite descriptions that have a different head from their antecedent (which Vieira and Poesio call *bridging* descriptions), they use synonymy/hypernymy/part-of and co-hyponym relations to successfully resolve 39% of all such bridging descriptions (from under 10% if just taking the closest noun phrase as an antecedent). Leaving aside precision errors, this shows that WordNet lacks coverage for many of the *bridging* relations found in the corpus; Work using distributional similarity as a ranking criterion [57] shows that this only results in appropriate antecedents for 23.6% of such definite descriptions, perhaps underlining the importance of targeting specific relations instead of using a general relatedness measure.

Markert and Nissim [44] contend that some of this stems from relations that are asserted purely in the text and not holding globally: In example 27 (from [45]), the text constructs a relation of *age* being a *risk factor* that we would not expect to find in any realistic ontology.

(27) You either believe Symour can do it again or you don't. Beside *the designer's age*, other risk factors for Mr. Cray's company include the Cray-3's [. . .] chip technology.

Markert and Nissim's solution then, is to use pattern-based text mining to uncover hyponym relations that are asserted in a text, even when they would not hold in principle or globally. They compare the use of WordNet with (i) pattern mining on the British National Corpus [10], but also (ii) using a search engine to query patterns on the World Wide Web. They show that it is possible to increase recall from 56.2% (for string matching only) to 64.9% using WordNet, 59.7% using the BNC, or 71.3% using Web search when constrining antecedents to match in number, with precision that still reaches 62.7% (BNC) or 71.3% (Web) for resolving a definite description that is known to be discourse-old.

Gasperin and Vieira [21] use a word similarity measure (from [22], very similar to Lin's [41] measure). In contrast to Poesio, Schulte im Walde and Brews [57] work, they do not resolve to the semantically closest noun, but instead build lists of globally most similar words (a so-called *distributional thesaurus*), and enable the resolution to antecedents that are in the most-similar list of the anaphoric definite, where the antecedent has the anaphoric definite in its most-similar list, or where the two lists overlap. Working on Portuguese data, Gasperin and Vieira find that they reach similar levels of resolution accuracy to the earlier results of Poesio, Schulte im Walde and Brews with a window-based association metric.

The pattern-based approach requires large corpora to achieve a reasonable recall: this is because patterns occur rarely in corpora. Accordingly, researchers in CL turned in the last years to the Web as a very large resource of linguistic data and developed a variety of knowledge acquisition methodologies (typically using weakly supervised techniques) to mine this large repository of text.

In a similar fashion, Poesio et al. [58] use a multilayer perceptron with features including simple graph distance in WordNet (indicating the number of nodes between the anaphor and the potential antecedent) and a feature based on the raw count of matches for a search engine query using a meronymy pattern. To express salience, Poesio et al. include the sentence distance to the anaphor, but also whether it is in first-mention position, or if any preceding mention of the entity had been in first-mention position.

Bunescu [9] proposes to use discourse-based patterns in conjunction with web queries to resolve bridging anaphora: To resolve an associate definite description to an antecedent, he embeds anaphor and antecedent noun phrases in a pattern *Y*. The *X verb*, where *verb* is subsequently filled with a list of auxiliary and modal verbs, and results are scored using pointwise mutual information. On a test set of associative bridging anaphora sampled from the Brown corpus section of the Penn Treebank, Bunescus approach reaches a precision of 53% at a recall of 22.7%. A very similar approach is presented by Garera and Yarowsky [20], who investigate the use of an unsupervised model to extracts hypernym relations from cooccurrence statistics for resolving definite nominals. The method aims at exploiting association metric scores to find likely categories for named entities: using the English Gigaword corpus as

source of textual data, they evaluate on a hand-selected sample and show that, when using the same corpus, their association measure can achieve greater recall than Hearst-style patterns.

Versley [82] tackles the question of more efficient combination of hand-annotated resources (such as GermaNet [36], a German wordnet) and unsupervised learning from corpora, more specifically the use of generic distributional similarity. He finds that for the problem of selecting an antecedent in the setting used by Markert and Nissim, or Garera and Yarowsky, syntax-based distributional similarity measures are more effective at improving recall when simply used in ranking (similar to Poesio, Schulte im Walde, and Brews approach using a window-based association measure) than when using it via the intermediate of a distributional thesaurus. However, these measures offer a similarly low precision as semantic classes (which can be computed with high accuracy using GermaNet and other hand-crafted resources, but only offer very limited information). As distributional measures such as that of Padó and Lapata [55] are not strictly limited to semantic classes, however, it is possible to fruitfully combine filters for distributional similarity, distance, and model-assigned semantic class to reap large improvements in precision at a small cost in recall (yielding 80% overall precision and 59% overall recall, against 70% precision and 64% overall recall for the unmodified distributional similarity and 67% overall precision and 62% overall recall for semantic classes only). In combination with GermaNet and Web-based pattern search, it is thus possible to find a coreferent antecedent for definite descriptions with 79% precision and 68% recall, or 73% overall F-measure. As these experiments (like those of Markert and Nissim, or Garera and Yarowsky) assume gold-standard information on the discourse-old/discourse-new distinction, the benefit of these recall-oriented resolution methods is likely to be less pronounced in a more realistic setting that takes into account discourse-new classification.

14.2.2 A Rush on Gold Mentions

Harabagiu et al. [26] make extensive use of WordNet, including non-taxonomic relations, for different coreference resolution subtasks in MUC-style coreference resolution. They go beyond synonymy and hypernymy and consider **more general paths in WordNet** that they find between anaphor-antecedent pairs found in the training data. To find candidate pairs, they filter out anaphoric expressions with an antecedent that can be found with knowledge-poor methods, such as string matching, appositions, name variation, or the most salient compatible antecedent. For the remaining anaphoric definite expressions, they look for anaphor-antecedent pairs that are related by at most five of the following relation types in the WordNet graph:

- SYNONYM, ISA/R-ISA and HAS-PART correspond to synonymy and hypernymy and meronymy relations.
- GLOSS/DEFINES connect a word in a synset to the word used to define it.

- IN-GLOSS/IN-DEFINITION connects an element of the synset with one of the first words in its definition.
- MORPHO-DERIVATION connects morphologically related words.
- COLLIDE-SENSE connects synsets of the same word (homonyms or polysemous senses).

Harabagiu et al. use three factors to measure the confidence of a WordNet path to predict a coreference relation. The first factor is a binary-valued flag that is set to 1 if another coreference chain contains mentions in the same nominal as the anaphor and the antecedent e.g. given *Massimos son* and *his bicycle*, if *son* and *his* have been previously found to be coreferent, the factor for the former pair is set to 1, else to 0.

The second factor prefers stronger relations where each WordNet relation type is assigned a weight ranging from 1.0 for SYNONYM over 0.9 for ISA and GLOSS down to 0.3 for IN-GLOSS). The weight is averaged over the relation types occurring in the path, with multiple occurrences of a relation weighted down by a factor corresponding to their number of occurrences. Additionally, the total number of different relations is used to weight down longer paths.

As an example, a path with one HASPART edge (weight 0.7) and two ISA edges (weight 0.9) would receive a weight of $\frac{1}{2} \cdot \left(\frac{0.7}{1} + \frac{0.9}{2}\right) \approx 0.57$, whereas a path with two ISA edges would receive a score of $\frac{1}{1} \cdot \frac{0.9}{2}$.

Finally, the last factor is a semantic measure inspired by the tf-idf weighting scheme and it is determined by considering the search space built when considering at most five combinations of the semantic relations defined above, starting from either of the synset a nominal can be mapped to. The overall confidence of a path is given by a weighted harmonic mean of the three factors. Confidence scores are then used to iteratively select the paths with the highest confidence as rules of the system.

By exploiting lexical knowledge from WordNet in a flexible way, Harabagiu et al.s proposal is able to achieve a very visible improvement in MUC F-measure from 72.3% to a 81.9%, albeit on gold-standard mentions together with the MUC F-measure, which favors low-precision, high-recall approaches. Later work by Luo et al. [43] points out that in this setting, putting all mentions of the test set into one coreference chain yields an impossibly high baseline of 88.2% F-measure (with 100% recall and 78.9% precision).

Poor evaluation and the ad-hoc-ness of their weighting functions aside, Harabagiu et al.'s work is noteworthy in that they use WordNet to derive a general distance measure, including the definitions contained in the glosses, which yield a markedly different information source from Poesio et al.s earlier approach (more focused on using the information in WordNet as it is and getting highly precise subsumption and synonymy predictions). They also use a global clustering-based model that can make use of more reliable decisions (e.g. for possessive pronouns) to influence other decisions (for the possessed NPs) where the coreference between the possessors provides additional information, something that would be nontrivial to incorporate into a machine learning approach.

Ponzetto and Strube: Relatedness in Wikipedia

Incorporating semantic knowledge into a machine-learning based system for coreference resolution – in this case WordNet, the use of Wikipedia's category hierarchy, and the use of Semantic Role Labeling – is the main goal of presented in Ponzetto and Strube [60],

Ponzetto and Strube use the category tree from Wikipedia as an unlabeled semantic network and compute semantic relatedness scores by means of taxonomy-based semantic distance measures previously developed for WordNet [65, 85, 38, 74].

Starting from the baseline system from Soon et al. [75], they extend it with different knowledge sources, including semantic distance scores computed from WordNet and Wikipedia, and present experiments on the ACE 2003 dataset. The authors find a large improvement in terms of recall on the broadcast news section (whereas the results on the newswire section are modest), with Wikipedia-based scores performing on par with WordNet. WordNet and Wikipedia features tend to consistently increase performance on common nouns. However, semantic relatedness is found not to always improve the performance on proper names, where features such as string matching and alias seem to suffice.

Semantic similarity computed from the Wikipedia taxonomy is evaluated extrinsically by Ponzetto [59], who use them as features of a supervised coreference resolver in the same way as the previously used semantic relatedness scores. The evaluation on the ACE-2 data show that using relatedness works better than computing paths along the ISA hierarchy. Semantic relatedness always yields better results than using similarity scores: but while this is a counterintuitive result, the author argues that this behaviour is an artifact of the annotations in ACE. The use of the Geo-political entities (GPE) class in the ACE data allows for coreferential links such as *Beijing ... China* and therefore mixes metonymy with coreference phenomena cf. the discussion in the introduction to this chapter and the link between Israel and its residents in the example 26. To generate these coreference links one needs indeed a more permissive notion of semantic compatibility, i.e. semantic relatedness. Using ISA relations only is (rightfully) expected to work better for data modeling coreference as identity only, which is the case for OntoNotes and most non-English corpora.

Ji et al.: Relation detection and coreference

An alternative to knowledge-lean approaches leveraging existing resources and unsupervised approaches extracting structured knowledge from unstructured textual resources is to learn semantic regularities directly from the same coreferentially annotated corpora used to train supervised coreference resolvers. Ji et al. [28] use heuristics integrate constraints from relations between mentions with a coreference resolver. The methodology consists of a two-stage approach where the probabilities output from a MaxEnt classifier are rescored by adding information about the semantic relations between the two candidate mentions. These relations are auto-

matically output by a relation tagger, which is trained on a corpus annotated with the semantic relations from the ACE 2004 relation ontology. Given a candidate pair 1.B and 2.B and the respective mentions 1.A and 2.A they are related to in the same document, they identify three lightweight rules to identify configurations informative of coreference:

1. If the relation between 1.A and 1.B is the same as the relation between 2.A and 2.B, and 1A and 2A dont corefer, then 1.B and 2.B are less likely to corefer.
2. If the relation between 1.A and 1.B is different from the relation between 2.A and 2.B and 1.A is coreferent with 2.A, then 1.B and 2.B are less likely to corefer.
3. If the relation between 1.A and 1.B is the same as the relation between 2.A and 2.B and 1.A is coreferent with 2.A, then 1.B and 2.B are more likely to corefer.

While Ji et al. argue that the second rule usually has high accuracy independently of the particular relation, the accuracy of the other two rules depends on the particular relation. For example, the chairman of a company, which has a EMPORG/Employ-Executive relation, may be more likely to remain the same chairman across the text than a spokesperson of that company, which is in the EMPORG/Employ-Staff relation to it.

Accordingly, the system retain only those rule instantiated with a specific ACE relation which have a precision of 70% or more, yielding 58 rule instances. For instances that still have lower precision, they try conjoining additional preconditions such as the absence of temporal modifiers such as current and former, high confidence for the original coreference decisions, substring matching and/or head matching. In this way, they can recover 24 additional reliable rules that consist of one of the weaker rules plus combinations of at most 3 of the additional restrictions.

They evaluate the system, trained on the ACE 2002 and ACE 2003 training corpora, on the ACE 2004 evaluation data and provide two types of evaluation: the first uses Vilain et als scoring scheme, but uses perfect mentions, whereas the second uses system mentions, but ignore in the evaluation any mention that is not both in the system and key response. Using these two evaluation methods, they get an improvement in F-measure of about 2% in every case. In the main text of the paper, Ji et al. report an improvement in F-measure from 80.1% to 82.4%, largely due to a large gain in recall. These numbers are relatively high due to the fact that Ji et al. use a relaxed evaluation setting disregarding spurious links. A strict evaluation on exact mentions is able instead to yield an improvement in F-measure from 62.8% to 64.2% on the newswire section of the ACE corpus.

Ng: Semantic classes and similarity

Ng [53] includes an ACE-specific semantic class feature that achieves superior results to Soon et al.s method using WordNet by looking at apposition relations between named entities and common nouns in a large corpus to find better fitting semantic classes than using WordNet alone. In addition, he uses a semantic similar-

ity feature similar to the one introduced by Gasperin et al. (indicating if one NP is among the 5 distributionally most-similar items of the other), and two features that are learnt from an held-out subset of the training data:

- a pattern-based feature, encoding the the span between mentions by means of a variety of patterns, e.g. as sequences of NP chunk tokens;
- an anaphoricity feature which encodes how often an NP is seen as a discourse-old noun phrase in the corpus;
- a coreferentiality feature modeling the probability that two noun phrases are coreferent, estimated by looking at pairs occurring in the corpus.

Training on the whole ACE-2 corpus, Ng is able to improve the MUC score from 62.0% on the ACE-2 merged test set to 64.5% using all the features except the pattern-based one.

Yang and Su: Selecting Extraction Patterns

Yang and Su [86] present an approach to select patterns as features for a supervised coreference resolver. Starting from coreferent pairs found in the training data such as Bill Clinton and President (or, due to the annotation scheme of the ACE corpora, Beijing and China, cf. the example 6), they extract patterns from Wikipedia where a pattern is defined as the context that occurs between the two mention candidates e.g. "(*Bill Clinton*) is elected (*President*)".

To select those patterns that identify coreferent pairs with a high precision, the method filters out in a first step those that extracts more non-coreferent pairs than coreferent ones in the training data. In a subsequent step, patterns are ranked based either on raw frequency or on a reliability score and the 100 top-ranking patterns are kept. In the case of the frequency-based approach, a feature is created for each pattern that indicates the frequency of that particular word pair with the pattern in the Wikipedia data.

For the other approaches, they calculate a reliability metric for each pattern (determined by summing the pointwise mutual information values between a pair of noun phrases and the pattern, over all coreferent pairs from the training data). The score for a given pattern and a pair of fillers is then determined as the value of the reliability of that pattern multiplied by the positive mutual information between positive mention pairs. Yang and Su apply these features in a coreference resolution system similar to the one described by Ng and Cardie [54] on the ACE-2 corpus. Using the reliability-based single relatedness feature for proper names (the setting they found to work best) results in an improvement from 64.9% F-measure to 67.1% on the newswire portion, 64.9% to 65.0% on the newspaper portion, and from 62.0% to 62.7% on the broadcast news part.

Daumé and Marcu: WordNet and Patterns

An integrated approach is presented in Daumé III and Marcu (2005), who use several classes of features. Besides including WordNet graph distance and WordNet information for preceding/following verbs (in an attempt to let the coreference resolver learn approximate selectional preferences in a supervised way), they also use name-nominal instance lists mined by Fleischman et al. from a large newspaper corpus [19], as well as similar data mined from a huge (138GB) web corpus [71]. They also used several large gazetteer lists of countries cities, islands, ports, provinces, states, airport locations and company names, as well as a list of group terms that may be referenced with a plural term.

14.3 Current Approaches

From the previous section, it should be clear that a large body of work exists that predates the recent shared-task evaluations in SemEval-2010 [72] as well as CoNLL-2011/2012 [63, 64], starting on small datasets created for the investigation of one particular problem, then ongoing with the MUC and ACE corpora; however, inconsistent evaluation practices make it somewhat difficult to compare approaches by different authors and/or to quantify the impact of techniques more precisely. This is compounded by the use of non-realistic settings using gold mentions that creates a picture of the usefulness of high-recall resolution techniques that does not correlate with performance in practice. Hence, the following investigation will try to gather a coherent picture of the problem based on work that uses a common corpus (usually the CoNLL version of the OntoNotes corpus) and on a standardized implementation of the evaluation metrics (as released in the CoNLL scorer).

14.3.1 Quantifying the Problem

If we want lexical/encyclopedic features to be effective, we should focus on the phenomena that have the most impact, possibly also on those where we can get the most precise description.

One of the most well-known papers comparing the behaviour of multiple coreference resolution systems is the one by Kummerfeld and Klein [35], who compares several open-source systems [39, 16, 6, 77, 83, 4, 66] as well as the participants of the CoNLL-11 shared task. In the context of this chapter, we see that their "*average of top ten systems*" data contains about as many extra as missing mentions for proper names, of which more than three quarters have matching text or matching head, whereas for nominals we see many more extra than missing text-matching mentions than missing ones; on the side of mentions without common text or com-

mon head, there are many more missing than extraneous ones. However, we need to look a bit further for a more detailed analysis.

The most detailed analysis of system coverage for coreference resolution on the OntoNotes/CoNLL dataset to this date has been performed by Martschat and Strube [47].[9] Their goal is to distinguish types of recall errors (i.e., links between mentions/entities that the coreference system should have made, but didn't make), and to this end they first transform the mention clusters of the original coreference dataset (where a set of mentions referring to one entity is annotated as belonging together) to a directed graph of antecedence relationships.

Based on a superset of directed edges linking all mentions to each other – Martschat and Strube assume that the introducing mention comes before the subsequent one, excluding cataphoric relationships – they give priority to edges representing true positive links where the given coreference system and the gold standard agree.

Beyond that, they apply different heuristics to weight the edges that are aimed to uncover sensible antecedence relationships:

- antecedence links from an *uninformative preceding mention* to more informative following one are avoided, in particular where the preceding mention is a pronoun and the following one is a non-pronoun, or where the preceding mention is a nominal one and the following one is a name.
- antecedence links are *weighted by distance*, such that the closest link is chosen among several that are otherwise admissible.

Like Kummerfeld, Martschat considers the output of different systems, including the Stanford Sieve resolver [39], his own system [46], IMSCoref (see [6], or the description in Chapter 8) as well as the BerkeleyCoref system (see [16] or the description in subsection 14.3.3). The Stanford system and Martschat's CoRT system are rule-based resolvers using heuristics for cannot-merge constraints and steps that create links among several mentions, whereas IMSCoref and BerkeleyCoref are based on machine learning classifiers (and consequently can use, e.g., a lexicalized feature model).

In his investigation, Martschat finds that the Stanford Sieve and IMSCoref make the most errors, whereas less of them occur with CoRT and BerkeleyCoref. He also investigates how of the recall errors are *common* to all the coreference systems: only half of the errors in coreference between names are common, while other categories show around 80% overlap in the errors that each system makes.

Coreference relations among name mentions, presumably the most interesting as not all systems agree on them, could already benefit from better approximate string matching (where clearly the difficulty in approximate string matching is to cover more cases *while maintaining good precision*).

Among these 475 cases of missed links between **two names**, about 154 have a complete string match, which means that either annotation inconsistencies or errors in mention extraction would be responsible for them (e.g., a non-match between

[9] Both the error analysis code and the code for Martschat's CoRT system are publically available at `https://github.com/smartschat/cort`

China and *China's*). A further 109 of these missing links have at least one token in common, pointing to approximate string matching that operated too cautiously. Then, about 104 cases remain where the two mentions share no token in common, which are in majority due to date matching, aliases (e.g., *Florida* and *the Sunshine State*), and acronyms, followed by a long tail of metonymy (e.g., using the capital city's name metonymously for a country), roles (such as *Al Gore*. . . *Vice President*), as well as names that are used with inconsistent spelling (especially foreign names).

Martschat found 371 links with a **nominal** having a **name antecedent**, which are often knowledge-dependent (for example, *Mr. Papandreou* being *the prime minister*). Of these, Martschat found that the ten most frequent heads make up 88 of the 371 errors.

Finally, there are a great many links **between nominals** that are missed. Of the 835 instances here, 174 are ones where the subsequent mention is an indefinite noun phrase, which one would normally exclude on linguistic grounds. 341 links in total are noun phrases with matching heads, which again drives home the fact that some problems in coreference resolution are difficult to do *with enough precision*.

Martschat also investigates a sample of 50 coreference links between two nominals with differing heads; of these 50 instances, 23 are hyponyms, 10 are synonyms. Comparing with the rest of these numbers, we see that "boring" problems such as name-nominal matches or same-head nominals, which were not necessarily the most prominent case for the phenomenon-oriented papers of the mid-2000s, are actually fairly important to achieve good coreference resolution performance overall.

Looking at the different systems individually, several facts become apparent; the first is that supervised learning systems can resolve a number of different-head links correctly when they are **frequently coreferent** in the training data (see also subsection 14.3.3 for a discussion of this). In particular, the Berkeley Coreference system recovers some links not found by Martschat's CoRT system. On the other hand, both the Stanford system and Martschat's system can resolve some cases that the learning-based systems miss by using more sophisticated alias heuristics, but conversely they miss some cases by performing overly strict modifier agreement checks in cases with matching substrings.

In Martschat's comparison, matching between two names is easiest, but the number of precision errors varies substantially between the systems (where the Berkeley-Coref system shows about 24% precision errors and the Stanford Sieve gives about 31% precision errors). Matching name/nominal links (in either direction) is the least precise category among those investigated. Among the systems investigated, CoRT is the most precise in finding links between two nominals.

14.3.2 Using lexical and Encyclopedic Resources

As one of the more modern papers using resources containing lexical and encyclopedic knowledge, let us look at the work of Rahman and Ng [67], who use the YAGO ontology, a pattern dictionary (linking named entities to probable words for their

semantic type), as well as some resources for representing verbs using FrameNet and PropBank

Rahman and Ng start with a cluster-ranking coreference resolver, to which they add features that would specifically help in resolving cases where world knowledge would be helpful.

Like some or most of the work in section 14.3.5 below, Rahman and Ng use features based on **YAGO**, including both aliases and hyperonymy-type links, which would allow to resolve "*Martha Stewart*" to the WordNet concept "*celebrity*" via the intermediary of "Television personalities" (the Wikipedia category that *Martha Steward* is found in, and linked to WordNet in YAGO's ontology).

For pronouns, where knowledge-poor systems can only use morphological agreement and salience/recency, governing verbs as well as the antecedent's governing verbs can give a better indication of antecedent plausibility than the pronoun alone.

Rahman and Ng, in this case, use features based on **FrameNet** and **PropBank** (mostly using PropBank for assigning semantic roles, which are then assumed to be consistent across the verbs of a FrameNet frame) and try to assess whether a pronoun and its (potential) antecedent would fit together.

Specifically, one feature considers whether both governing verbs could be part of the same FrameNet frame (yielding a value of yes, if they could be, no, if the possible frames fo not overlap, or the information that one or both verbs are not part of FrameNet).

The other feature for Rahman and Ng's way of integrating frame information into the coreference resolver is to look at the semantic roles assigned by the PropBank labeler, limiting the consideration to ARG0 (proto-subject) and ARG1 (proto-object) and leaving out all other roles to avoid sparse data. The feature using this information then indicates whether anaphor and antecedent are both ARG1, both ARG1, show a transition (ARG1 to ARG0, or ARG0 to ARG1), or that one or both of the noun phrases have a different role than ARG0 or ARG1.

Rahman and Ng combine both of these role-related features, yielding 15 binary-valued features. To apply these to cluster ranking, they add a feature whenever it is true for at least one of the mentions in the candidate's cluster.

Besides features based on specific resources such as YAGO or FrameNet, Rahman and Ng also use features that are based on a simpler, more direct representation of the mentions or their context, using, e.g., the heads themselves as features.

In their **noun pairs** feature, Rahman and Ng represent the mention and corresponding antecedent candidate as an ordered pair of heads, replacing named entities by a more general representation (either the label from the named entity recognizer itself, in the case when one is a common noun and one is a name, or the concatenation of both NE classes when both are names, replaced by "[*class*]-SAME" and "[*class*]-SUBSAME" features whenever two mentions have the same named entity class and either their strings match or there is an overlap in a subset of the tokens.

To improve the generalization capability for the fully lexicalized features, Rahman and Ng replace 10% of the common nouns in training by a special UNSEEN label, meaning that the fully lexicalized features are not used, whereas "UNSEEN-

SAME" and "UNSEEN-DIFFERENT" features are used based on whether the anaphor and the antecedent-candadate show string identity or not.

To extend coverage beyond that of FrameNet, Rahman and Ng also include a feature based on the PropBank roles together with just verb lemmas (instead of frames), which uses the **verb pair** together with the respective roles.

To further increase the coverage for noun phrase-name pairs, Rahman and Ng use existing repositories of NP pairs extracted through patterns [19, 53], covering, for example pairs such as "*Eastern Airlines*" and "*the carrier*", for a total of slightly over one million NP pairs. They use a binary-valued feature indicating whether the anaphor and any mention in the candidate cluster can be found in the extracted database.

In their evaluation, Rahman and Ng build the additional world knowledge features into a state-of-the art coreference resolver based on cluster ranking, and find that all of the proposed features give improvements (sometimes small, sometimes rather visible).

Among the features aimed at non-pronouns, their *YAGO-Types* feature (aimed at hyperonymy-like relations within YAGO) and the non-resource-using *word pairs* features perform at about the same level, with the resource-based YAGO-Types feature sometimes reaching better precision, and both features reaching better precision as well as recall than the pattern-based *Appositives* feature or the *YAGO-Means* feature, which only targets aliases and synonymy.

Among the verb-based features, which should make a difference mainly for pronouns and other not-as-informative noun phrases, Rahman and Ng report that the *verb pairs* feature performs better than the *FrameNet* feature on both precision and recall, yielding a consistent advantage over all evaluation measures, corpora, and system variants (mention-pair vs. cluster ranking).

Altogether, Rahman and Ng get a very large improvement of about 4% from the combination of all features, across different corpora (ACE and OntoNotes) and evaluation metrics (Bcubed and CEAF). In particular, this is a difference about as large as that between cluster-ranking and the simpler mention pair model. They also show that, among the low-hanging fruits in terms of features that can be used with just a large training corpus and no additional lexical resources, the largest gains can be achieved using the noun-pairs features, followed by those using pairs of adjective phrases and then pairs of verbs.

Bansal and Klein [2] present another approach to use external resources – in this case, patterns extracted from Google's Web n-grams dataset covering both the case of name-noun patterns used in previous work but also more general co-occurrence statistics based on this data.

Bansal and Klein start from a mention-pair model similar to the Reconcile system of [76], but using a decision tree learner (which gave more precise results in their study, with a relatively small cost in terms of recall), and add additional features to capture more information on *general lexical affinities* (i.e., relatedness expressed through co-occurrence), *lexical relations* (i.e., relations expressed through patterns such as those used by Rahman and Ng, earlier this section), *similarity of entity-based context* (typical verbs or adjectives co-occurring in pre-defined patterns), as well as

matches of the soft clustering from an existing dataset [42], as well as statistics about plausible fillers for a *pronoun context*.

In the simplest case, **co-occurrence of the head words** within a 5-gram window is counted, and normalized by the counts of each individual head word by itself, and using a binned log-value of this ratio as a feature (except for a normalizing overall count, this statistic is very similar to the *pointwise mutual information* statistic, and indicates the degree of first-order relatedness of the two items).

For the **patterns co-occurrence** feature, Bansal and Klein use Hearst-like co-occurrence patterns indicating isA-relations. In particular, they use patterns of the form

- h_1 BE DET? h_2
- h_1 (and|or) (other|the other|another) h_2
- h_1 (other than|such as|, including|, especially) DET? h_2
- h_1 of (the|all) h_2

(Where BE corresponds to the forms *is*/*are*/*was*/*where*, and DET corresponds to the forms *a*/*an*/*the*).

They use these patterns in the form of a quantized normalized count for all patterns together, using a binning strategy similar to the general co-occurrence feature.

For the **entity-based context** feature, Bansal and Klein look for copula or passive constructions linking one head to a descriptive word using the "*h* BE DET? *y*" pattern (finding not only nouns such as *head* for president, but also predicates such as *elected* or *responsible*), using the list of the frequent 30 matches to construct a descriptor for a head word, and turn these descriptor list for one feature that indicates whether there is a match within the top k items on both lists (leaving the k as a tuning parameter), and one whether both lists have the same dominant part of speech (for one of adjectives, nouns, adverbs or verbs), or *no-match* otherwise.

For the **cluster-based** similarity, Bansal and Klein look for the topmost-ranked cluster-id that is common to the (20 most salient) clusters for two heads, and use the rank sum as a feature in binned form.

Finally, their **pronoun context** feature is most useful for pronoun anaphors in that it also looks at the context of the mention: given a context "(pronoun) *r r′*", Bansal and Klein try to estimate how likely it is that the pronoun has an antecedent "h_1" that would fit this context by considering co-occurrences of h_1 with the first right context word either directly neighbouring (R1) or with a gap in-between (R1gap), or with both right context words directly neighbouring (R2).

As an example, if "*his*" in "*his* victory" is considered with a candidate "*Bush*" as h_1, the normalized count

$$\frac{\text{count}(\text{"}\textit{Bush's}\star\textit{victory}\text{"})}{\text{count}(\text{"}\star\textit{'s victory}\text{"})\,\text{count}(\text{"}\textit{Bush}\text{"})}$$

would be calculated for the R1gap feature. Counts for the different variations are binned and used as separate features.

In their experiments, which use the MUC metric and the B^3all variant of the Bcubed evaluation metric, they find that they get a large improvement (about 4%)

from switching from a linear classifier (averaged perceptron) to a decision tree learner, and another gain of about 1.1-1.8% for all the web-based features together, with the largest improvement coming through the pronoun context feature (+0.6% MUC, +0.4% B3all), the Hearst patterns (+0.2% MUC, +0.4% B3all) as well as the headword similarity through context features (+0.4% MUC, +0.1% B3all).

Bansal and Klein report that, in comparison to their baseline system, the new system is more successful at resolving many name-noun coreference links such as finding "*the EPA*" as the antecedent of "*the agency*" or "*Barry Bonds*" as antecedent of "*the best baseball player*", and that strong general co-occurrence, as well as the absence of cluster-match and/or Hearst pattern matches, as well as the R2 pronoun context feature are strongly discriminative for the decision tree.

14.3.3 Lexicalized Modeling of Coreference

Two mostly recent innovations – namely the predominance of learning models that (like maximum entropy, support vector machines, or structured learning counterparts of these, but unlike decision trees) support high-dimensional feature spaces on the one hand, and the availability of relatively large coreferentially annotated corpora, makes the use of **lexicalized features** such as the *noun pair* or *verb pair* features used by Rahman and Ng in the previous section both possible and (as it turns out) quite attractive.

As examples of systems that can use more detailed information in a purely supervised learning scenario that does not involve external resources, we will have a look at two relatively modern systems (Durrett and Klein's *Easy Victories* system [16] as well as Bjrkelund's *HOTCoref* system [7], also described in Section 14.3.6 and Chapter 8, respectively). Durrett and Klein use the term *Easy Victories* to describe the fact that cluster ranking with an online learning approach (as used by Durrett and Klein, or by Bjrkelund's HOTCoref) allows it to make use of lexicalized features quite easily, and that these lexicalized features offer a straightforward way to improve the resolution of more difficult cases whenever the relevant information occurs frequently enough in a large corpus.

To start with the learning and inference part, Durrett and Klein use a mention-synchronous resolver similar to that of Luo et al. [43] or Daumé and Marcu [15], applying loss-scaled online updates whenever a wrong resolution decision in training occurs; this avoids the training set balance that plagued classical mention-pair systems such as [75] or [54]. By assigning different (local) losses to different types of wrong decisions (*false negatives*, *wrong links*, or *false anaphors*), Durrett and Klein can model the decisions for structure-building in coreference in a suitable way.

Durrett and Klein combine their individual features from general information (type of both mentions - names, nominals, or various types of pronouns), more specific features on the current mention, or more specific features on the antecedent candidate – in particular, *head word*, as well as *first*, *last*, *preceding* and *following*

words as well as the length of that mention – or alternatively, features that fire based on the pair, namely string match of the heads or of the complete string, and distance in terms of sentences or mentions (capped at 10).

Durrett and Klein show that their surface lexicalized features successfully cover distinctions such as definiteness, number/gender/person matching for pronouns, as well as some information on the grammatical role of a mention. Using specialized features for these criteria, as well as WordNet-based hypernymy and synonymy, an existing number and gender dataset (see Chapter 12), as well as named entity types and rough clusters of nominal heads and verbal roles, then, gives an additional gain leading from a CoNLL metric score of 60.06 to 60.42, which is substantially less impressive than the gains from the "generic" lexicalized features.

The reason for this rather small gain from more semantic features, according to Durrett and Klein (all while they achieve a substantially larger gain – from 75.08 to 76.68 – in the non-realistic setting with gold mentions) is the insufficient precision of the cues for non-same-head mentions.

For their *final* system, Durrett and Klein combine the surface features with additional features that do not necessarily target semantics but are proven to be helpful, in particular whether two mentions are nested, a dependencies feature including the parent and grandparent POS tags and arc direction, as well as a speaker identification feature, on top of the aforementioned gender/number data. Using all these features, Durrett and Klein's *final* system reaches a CoNLL score of 61.58, indicating that their framework, together with a simple homogeneous set of features together with a small set of (non-semantic!) specialized features yields relatively high performance.

14.3.4 Knowledge-based Alias Resolution

An approach that is conceptually different is used by Recasens, Can and Jurafsky [73], who use a corpus of comparable documents to extract **aliases** (i.e. pairs of nouns or names that uniquely refer to the same thing, but which may not be similar on the surface level), such as *Google* and *the Mountain View search company*.

Recasens et al. start by downloading clusters of documents about the same events from the *Techmeme* news aggregation site, obtaining 25k story clusters totalling about 160 million words. Among the documents pertaining to the same story, they rank the verbs according to their tf-idf score (excluding light verbs and reporting verbs), to gather comparable mentions of events such as *Google crawls web pages* and *The search giant crawls sites*, and assuming that the subjects and the objects each are referring to the same thing unless one of a list of *filtering* criteria is fulfilled, which are added to exclude some spurious pairs (raising the precision from 53% to 74%): If both mentions are named entities, if at least one of them is a number or temporal NE, or the mention of the verb has a negator, the pair is filtered out.

In a second step, Recasens et al. remove determiners as well as clausal modification, yielding a core consisting of the head noun or name, adjectival or genitive

premodifiers, and PP postmodifiers. They also generalize non-head named entities to their types, allowing to link "*the leadership change*" to "⟨*Person*⟩*'s departure*". Recasens et al. report that their method finds not only synonymy and instance relations such as *change* versus *update*, but also metonymic cases such as *content* versus *photo*, or *government* versus *chairman*.

Recasens et al. incorporate the extracted dictionaries in the rule-based *dcoref* resolver, and evaluate their system using gold mentions, a setting with which they achieve a gain of 0.7 percent F1 score for the CoNLL metric.

14.3.5 Combining NE Linking and Coreference

While it has been a long-term goal of some work (e.g. Ponzetto and Strube, 60) to include knowledge based on names in coreference resolution, both the ambiguity resolution of names and the mechanisms to use knowledge about entities have improved in recent work: In particular, the current state of the art includes newer resources such as the YAGO ontology, as well as annotated data on general-purpose linking of names to Wikipedia, either in the form of data generated from Wikipedia itself or in the shared tasks of the Text Analytics Conference [29, 30]. In particular, neither Rahman and Ng nor Ponzetto and Strube make use of such a dataset or the taggers derived from that data to disambiguate mentions of entities.

Uryupina et al.: Name disambiguation and YAGO relations

Among the first to use name disambiguation in a coreference systems were Uryupina et al. [80], who use a name disambiguation approach based on training data derived from Wikipedia based on a kernel capturing gappy n-gram, single-word and latent semantic information, and subsequently using YAGO MEANS and TYPE relations as features.

To improve the precision of their YAGO-based features, Uryupina et al. use several filters that exclude cases that often yield false positives:

The first one is **discourse-new detection** for potential hyperonyms: If a candidate for a hyperonym has a modifier indicating a non-anaphoric mention, such as *any other country* in

(28) [India]'s advantage, it simply has more skilled, English-speaking programmers than [any other country] outside the U.S.

where Uryupina et al. suppress the indication of a YAGO TYPE relation whenever the determiner of the subsequent noun phrase is incompatible with a coreference relation. The second covers **too common hypernyms**, which are terms at a very general level of the taxonomy such as *group* or *part* which frequently occur in false positives, based on a manual analysis. A third filter prohibits the feature from firing

for nominal-name links (as opposed to name-nominal ones) since these are often part of wrong or inconsistent coreference decisions.

Uryupina et al. test their approach on the ACE-02 corpus, against two baselines: one using Soon et al's [75] features, and one additionally using a WIKI-ALIAS feature that detects mentions linking to the same concept in Wikipedia.

In Uryupina et al.'s work, improvements were not uniform across the sections of the ACE data, but they show that the disambiguation, as well as the features based on YAGO relations and the additional filters can improve performance quite a bit – 1.6% CEAF and 2% MUC for gold mentions on BNEWS, or 1.4% and 1.8% over the baseline with WIKI-ALIAS feature, whereas the improvements for the other parts of the ACE corpus are less drastic.

While Uryupina et al.'s improvements are not uniform across all parts of the ACE-02 corpus, it should be noted that they show both the use of disambiguating and linking named entities to external ontologies as well as the usefulness of having interpretable structure rather than just using super- and subcategories within Wikipedia.

Ratinov and Roth: Relatedness through attributes

A number of subsequent approaches have used more elaborate learning approaches together with gold mentions, such as Ratinov et al. [69], who extract **attributes** from Wikipedia pages (such as *Redmond* for *Microsoft*), which they then use to further the recall in a system based on a hybrid of Lee et al's Sieve and a more standard mention-pair model, using the GLOW named entity linker of [70].

Ratinov and Roth use the attributes both for name-name and name-noun candidates, where they have different roles: two names of the same kind are very likely not to corefern, whereas a name and a matching noun often indicate an anaphoric relation.

Hajishirzi et al.: Comparibility/Incomparibility

A similar idea underlies the NECo system of Hajishirzi et al. [25]: They start from the Stanford Sieve system (see Chapter 3 for a discussion), but incorporate both the detected spans and the suggested linked entities in the mention clusters, and add to the sieve steps of the Stanford system two additional ones making use of the NEL information: one that is akin to the YAGO-MEANS relation which merges two coreference clusters if they link to the same entitiy, and one more akin to the YAGO-TYPE relation which merges common noun mentions with antecedents that have this noun as an attribute in Freebase (which covers descriptions such as *Donald Tsang* being a *president*, or *Disneyland* being a *park*).

Hjishirzi et al. use an ensemble from the named entity linkers GLOW [70] and WikipediaMiner [51] to extract high-confidence named entity link candidates, where

mention spans from both the coreference system and the from the NEL are merged and those duplicates that differ only by a stop word.

- an *exact link* l_m if the entire span excluding stop words links to a known entity.
- a *head link* h_m if the head matches a known mention (e.g., *President Clinton* to *Bill Clinton* because of the head word *Clinton*)

Clusters correspondingly receive an exact link l_c and a head link h_c based on their most prominent mention (the *exemplar* in the parlance of the Stanford Sieve). Clusters are regarded as incompatible if they have incompatible (i.e., non-null and different) exact or head links.

Additionally, the algorithm keeps track of lists of related entities:

- a list L_m of all entities with a direct link to the span's entity (including, for example, *Alabama* for *Bill Clinton*, or, for *The governor of Alaska Sarah Palin* references to *List of governors of Alaska*, *Alaska* and *Sarah Palin*.
- a list L'_m that additionally includes entities linked to sub-phrases
- A list L_c for each cluster that contains *all* linked entities found in a cluster

To be considered plausible merge candidates, two clusters must be *related* to each other in FreeBase; when merging, not only are the standard attributes of Sieve mentions are merged but also the union of the clusters' attributes is taken.

country	president	city	area
company	starte	region	location
place	agenc	power	unit
body	market	park	province
manager	organization	owner	trial
site	prosecutor	attorney	county
senator	stadium	network	building
attraction	government	department	person
origin	plant	airport	kingdom
capital	operation	author	period
nominee	candidate	film	venue

Table 14.1 The most commonly used fine-grained attributes from Freebase and Wikipedia (out of over 500 total attributes) in [25].

Hajizhirzi et al. compare their system with the Standford system on both the ACE-2004 and CoNLL-2011 datasets, including a fully automatic setting using system mentions and automatic named entity linking. They can achieve significant increases in the MUC score and slight increases on the B^3 score, and achieve larger improvements on the ACE-2004 dataset with respect to the Stanford Sieve system. They also compare the results on ACE2004-NWIRE to a version that omits the non-linking constraints between mentions that have incompatible NE links, showing that

a substantial part of the performance gains come from enforcing *in*compatibility between mentions sharing a common word (such as *Staples* the company and *Todd Staples* the politician).

14.3.6 Joint models of NEL and Coreference

Also on gold mentions, Zheng et al. [87] extend the idea of Ratinov and Roth to use *Dynamic Linking*, meaning that the coreference decisions of a mention-pair model can influence the subsequent decisions of named entity linking, using a reranking of the candidates. To do this, they use an existing named entity linking system [14], but keep (ambiguous) lists of candidate links, which they subsequently merge and rerank during inference.

A system with tighter integration between named entity recognition, entity linking and coreference in a **factor graph model for coreference and linking** was presented by Durrett and Klein [17], who model the decision in these tasks and the interactions across levels in a factor graph, using iterations of belief propagation in a pruned factor graph for inference in that model.

The intuition behind it all is simple: named entity recognition on ambiguous instances (e.g. shortened names) can profit from at least some kind of coreference information, and similarly can profit from Wikipedia knowledge, and similarly coreference can profit from better named entity information.

Features

The *coreference* variables (a_1, a_2 in Figure 14.1) in the model indicate, for each mention, whether it is new or should be resolved to some (previous-mention) antecedent, with the feature set of Durrett and Klein's earler lexicalized mention-ranking approach [16].

Named entity variables (t_1, t_2) specify, for each span, what semantic type (if any) they are, and consequently also describe the semantic type of their surrounding mention. They use state-of-the-art token-based features such as those used in earlier work on NER [68], including word clusters.

Finally, *entity linking* factors (e_1, e_2) link one mention to a particular Wikipedia title (or none); these are based on a query string that is supposed to be a latent variable (i.e., unobserved both in training and in testing) specifying which part of the mention string is queried for, as, e.g. *Chief Executive Michael Dell* has not been hyperlinked on Wikipedia whereas *Michael Dell* does. In addition, the relation between the "official" page title and the related text portion of the mention is modeled through a *query* variable (q_1, q_2).

Interaction factors

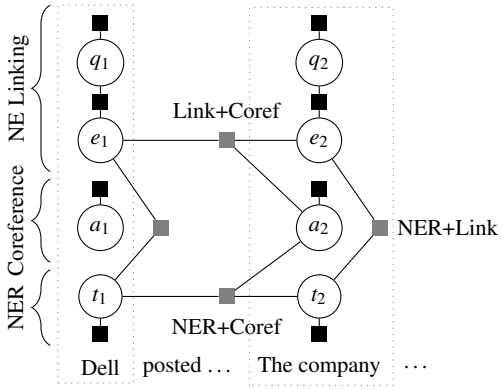

Fig. 14.1 Interaction between factors on different levels (from [17]): Consistency between Linking and NER is encouraged by NER+Link factors (e_1/t_1, e_2/t_2), Consistency between coreferent Link and NER mentions is ensured by Link+Coref ($e_1/e_2/a_2$) and NER+Coref ($t_1/t_2/a_2$) factors.

Durrett and Klein then link the variables on each level by cross-task interaction factors that use information across different tasks:

- The simplest factors concern *entity linking and NER*, in particular evidence such as categories (e.g. *American financiers*) or infobox types (e.g. *Person*, *Company*), as well as the predicate occurring in a copula clause in the first sentence of a Wikipedia article.
- Other factors enforce consistency between *Coreference* and *NER*: in particular, the pair of semantic types for the current and antecedent mention, and combinations of semantic type and head of the current mention, and the semantic type and head of the antecedent mention containing exactly one lexicalized item.
- Finally, factors joining *Coreference and Entity Linking* have a similar structure to those between coreference and NER, but use indicators of relatedness (same title, shared links, or links to each other)

Inference

To perform efficient inference with the joint factors, a first pass is carried out using only the within-level factors (which are substantially less complicated), to *prune* the possible variable settings, reducing both the size of the problem and the number of combinations that has to be considered for the joint factors.

Durrett and Klein use *Minimum Bayes Risk* decoding for the inference, meaning that they compute marginals over each variable (specifying how likely e.g. a given antecedence relation is) and compose the solution from most likely variable settings,

instead of simply using the single highest-scoring solution. This kind of inference can be advantageous to provide some isolation between the levels when it comes to uncertain results, at the potential disadvantage that the solution returned by this kind of inference may not always respect the consistency of multiple variables against each other.

Results

Durrett and Klein test their approach on a different test set for each task; for coreference in particular, they use the CoNLL-2012 and ACE-2005 test sets for the final evaluation of the coreference resolution performance of their system and compare their 2013 system (which does global coreference resolution based on a pipeline approach involving other components) to a version of their system without cross-task factors, and the complete, joint version of their system.

While the *independent* version of their system, presumably identical in approach to the 2013 version but with the better preprocessing that is part of the new system, Durrett and Klein achieve a score of 61.23 on the CoNLL-2012 test set, about 0.9 higher than the 2013 version. On top of this very competitive system, the interaction factors and joint decoding yields a further 0.5 improvement to 61.71, slightly higher than Bjrkelund's HOTCoref system (see [7] or Chapter 8).

14.4 Discussion and Outlook

In this chapter, we have discussed some of the issues around lexical and encyclopedic knowledge in more detail: from the problems that should be solved, to the lexical and encyclopedic knowledge that is available outside of annotated corpora to help in the decision, to the ways you can exploit both lexicalized models in large corpora and the incorporation of entity linking into corpora.

We saw that part of the confusion around using encyclopedic knowledge were works that perform evaluation in *non-realistic settings*, such as using gold mentions (i.e. using only mentions that are part of a coreference set in the gold standard, substantially simplifying the task) including early work by Harabagiu et al. [26]. Even more recent work such as Ponzetto and Strube [60] or even Ratinov and Roth [69] and Zheng et al. [87] that shows large gains from complex features does so in great part because these non-realistic settings are more forgiving of recall-heavy resolution strategies that sacrifice precision in order to find more coreference links.

Conversely, in realistic settings, where the loss in precision would be amplified by the additional (non-gold) mentions present, it is substantially harder to achieve gains by incorporate lexical and encyclopedic knowledge, but still possible and necessary, as demonstrated by most of the work discussed in the rest of this chapter, which uses more cautios techniques to achieve more modest (but practically relevant) performance gains.

According to the resources that were or are available, we see that there is a significant shift towards supervised *lexicalized* models for coreference that is fueled by the general availabilty of large corpora such as TBa-D/Z or OntoNotes; in the direct comparison of Rahman and Ng, we see that lexicalized features are low-hanging fruits in the sense that they achieve larger improvements than unsupervised learning (in part because supervised learning yields task-specific information while unsupervised learning only yields more general, and often more noisy, distinctions).

To use encyclopedic knowledge fully, Uryupina et al. and subsequent research by Ratinov and Roth, Hajishirzhi and Zilles *inter alia* shows that it is beneficial to use named entity linking, or disambiguation of entity mentions, to make full benefit of the information in Wikipedia, and that, contrary to the more extensive relationships sought after by the early work of Ponzetto and Strube, synonymy-like alias relations (YAGO-MEANS) and hyperonymy-like instance relation (YAGO-TYPE) are key to achieving access to precise, reliable lexical and encyclopedic information.

References

[1] Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada, pp 86–90

[2] Bansal M, Klein D (2012) Coreference semantics from web features. In: Proceedings of the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju Island, Korea, pp 839–398

[3] Bean D, Riloff E (2004) Unsupervised learning of contextual role knowledge for coreference resolution. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Boston, MA, pp 297–304

[4] Bengtson E, Roth D (2008) Understanding the value of features for coreference resolution. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, pp 294–303

[5] Berland M, Charniak E (1999) Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), College Park, MD, pp 57–64

[6] Björkelund A, Farkas R (2012) Data-driven multilingual coreference resolution using resolver stacking. In: Joint Conference on EMNLP and CoNLL – Shared Task, Jeju Island, Korea, pp 49–55

[7] Björkelund A, Kuhn J (2014) Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, MD, pp 47–57

[8] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: A collaboratively created graph database for structuring human knowledge. In: Pro-

ceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '08, pp 1247–1250, DOI 10.1145/1376616.1376746

[9] Bunescu R (2003) Associative anaphora resolution: A web-based approach. In: Proceedings of the EACL 2003 Workshop on the Computational Treatment of Anaphora, Budapest, Hungary, pp 47–52

[10] Burnard L (ed) (1995) Users Reference Guide British National Corpus Version 1.0. Oxford University Computing Service

[11] Carter DM (1985) Common sense inference in a focus-guided anaphor resolver. Journal of Semantics 4:237–246

[12] Charniak E (1972) Toward a model of children's story comprehension. PhD thesis, MIT Computer Science and Artificial Intelligence Lab (CSAIL)

[13] Dagan I, Justeson J, Lappin S, Leass H, Ribak A (1995) Syntax and lexical statistics in anaphora resolution. Applied Artificial Intelligence 9:633–644

[14] Dalton J, Dietz L (2013) A neighborhood relevance model for entity linking. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Lisbon, Portugal, pp 149–156

[15] Daumé III H, Marcu D (2005) A large-scale exploration of effective global features for a joint entity detection and tracking model. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada, pp 97–104

[16] Durrett G, Klein D (2013) Easy victories and uphill battles in coreference resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, WA, pp 1971–1982

[17] Durrett G, Klein D (2014) A joint model for entity analysis: Coreference, typing and linking. Transactions of the Association for Computational Linguistics 2:477–490

[18] Fernandes ER, dos Santos CN, Milidiu RL (2014) Latent trees for coreference resolution. Computational Linguistics 40(4):801–835

[19] Fleischman M, Hovy E, Echihabi A (2003) Offline strategies for online question answering: Answering questions before they are asked. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, Japan, pp 1–7

[20] Garera N, Yarowsky D (2006) Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), New York City, NY, pp 37–44

[21] Gasperin C, Vieira R (2004) Using word similarity lists for resolving indirect anaphora. In: ACL'04 workshop on reference resolution and its applications, Barcelona, Spain, pp 40–46

[22] Gasperin C, Gamallo P, Augustini A, Lopes G, de Lima V (2001) Using syntactic contexts for measuring word similarity. In: Proceedings of the ESSLLI 2001 Workshop on Knowledge Acquisition and Categorization, Helsinki, Finland, pp 18–23

[23] Ge N, Hale J, Charniak E (1998) A statistical approach to anaphora resolution. In: Proceedings of the Sixth Workshop on Very Large Corpora (WVLC/EMNLP 1998), Montreal, Canada, pp 161–171
[24] Giles J (2005) Internet encyclopedias go head to head. Nature 7070:900–901
[25] Hajishirzi H, Zilles L, Weld DS, Zettlemoyer L (2013) Joint coreference resolution and named-entity linking with mult-pass sieves. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, WA, pp 289–299
[26] Harabagiu S, Bunescu R, Maiorano S (2001) Text and knowledge mining for coreference resolution. In: Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001), Pittsburgh, PA, pp 55–62
[27] Hearst M (1992) Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING 92), Nantes, France, pp 539–545
[28] Ji H, Westbrook D, Grishman R (2005) Using semantic relations to refine coreference decisions. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), Prague, Czech Republic, pp 17–24
[29] Ji H, Grishman R, Dang HT, Griffin K, Ellis J (2010) Overview of the TAC 2010 knowledge base population track. In: Text Analytics Conference (TAC 2010)
[30] Ji H, Nothman J, Hachey B (2014) Overview of TAC-KBP2014 entity discovery and linking tasks. In: Proc. Text Analytics Conference (TAC 2014)
[31] Kameyama M (1997) Recognizing referential links: an information extraction prespective. In: ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, pp 46–53
[32] Kehler A, Appelt D, Taylor L, Simma A (2004) The (non)utility of predicate-argument frequencies for pronoun interpretation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), pp 289–296
[33] Kipper K, Dang HT, Palmer M (2000) Class-based construction of a verb lexicon. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI 2000), Austin, TX, pp 691–696
[34] Klebanov B, Wiemer-Hastings PM (2002) Using LSA for pronominal anaphora resolution. In: Proeedings of the Computational Linguistics and Intelligent Text Processing, Third International Conference, (CICLing 2002), Mexico City, Mexico, pp 197–199
[35] Kummerfeld JK, Klein D (2013) Error-driven analysis of challenges in coreference resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, WA, pp 265–277
[36] Kunze C, Lemnitzer L (2002) GermaNet – representation, visualization, application. In: Proceedings of LREC 2002, Las Palmas, Spain

[37] Landauer TK, Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition. Psychological Review 104(2):211–240

[38] Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed) WordNet, An Electronic Lexical Database, MIT Press, Cambridge, MA, pp 265–283

[39] Lee H, Chang A, Peirsman Y, Chambers N, Surdeanu M, Jurafsky D (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4):885–916

[40] Lin D (1995) University of Manitoba: Description of the PIE system used for MUC-6. In: Proceedings of the 6th Message Understanding Conference (MUC-6), Columbia, MD, pp 113–126

[41] Lin D (1998) Automatic retrieval and clustering of similar words. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (CoLing-ACL 1998), Montreal, Canada, pp 768–774

[42] Lin D, Church K, Ji H, Sekine S, Yarowsky D, Bergsma S, Patil K, Pitler E, Lathbury R, Rao V, Dalwani K, Narsale S (2010) New tools for web-scale n-grams. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta

[43] Luo X, Ittycheriah A, Jing H, Kambhatla N, Roukos S (2004) A mention-synchronous coreference resolution algorithm based on the Bell tree. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, pp 135–142

[44] Markert K, Nissim M (2005) Comparing knowledge sources for nominal anaphora resolution. Computational Linguistics 31(3):367–402

[45] Markert K, Nissim M, Modjeska NN (2003) Using the web for nominal anaphora resolution. In: Proceedings of the 2003 EACL Workshop on the Computational Treatment of Anaphora

[46] Martschat S (2013) Multigraph clustering for unsupervised coreference resolution. In: Proceedings of the ACL Student Research Workshop

[47] Martschat S, Strube M (2014) Recall error analysis for coreference resolution. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, pp 2070–2081

[48] Mendes PN, Jakob M, Bizer C (2012) DBpedia: A multilingual cross-domain knowledge base. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turke, pp 1813–1817

[49] Miller GA, Fellbaum C (1991) Semantic networks of English. Cognition 41:197–229

[50] Miller GA, Hristea F (2006) WordNet nouns: Classes and instances. Computational Linguistics 32(1):1–3, DOI 10.1162/coli.2006.32.1.1

[51] Milne D, Witten IH (2008) Learning to link with Wikipedia. In: Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), Napa Valley, CA, pp 509–518

[52] Nedoluzhko A, Mírovyský J (2013) How dependency trees and tectogrammatics help annotating coreference and bridging relations in Prague Dependency Treebank. In: Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pp 244–251
[53] Ng V (2007) Shallow semantics for coreference resolution. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, India, pp 1689–1694
[54] Ng V, Cardie C (2002) Improving machine learning approaches to coreference resolution. In: 40th Annual Meeting of the Asssociation for Computational Linguistics, Philadelphia, PA, pp 104–111
[55] Padó S, Lapata M (2007) Dependency-based construction of semantic space models. Computational Linguistics 33(2):161–199
[56] Poesio M, Vieira R, Teufel S (1997) Resolving bridging descriptions in unrestricted text. In: ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts, Madrid, Spain, pp 1–6
[57] Poesio M, Schulte im Walde S, Brew C (1998) Lexical clustering and definite description interpretation. In: AAAI Spring Symposium on Learning for Discourse, pp 82–89
[58] Poesio M, Mehta R, Maroudas A, Hitzeman J (2004) Learning to resolve bridging references. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004), Barcelona, Spain, pp 143–150
[59] Ponzetto SP (2010) knowledge Acquisition from a Collaboratively Generated Encyclopedia, Dissertations in Artificial Intelligence, vol 327. IOS Press, Amsterdam
[60] Ponzetto SP, Strube M (2006) Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006), New York, NY, pp 192–199
[61] Ponzetto SP, Strube M (2007) Deriving a large-scale taxonomy from wikipedia. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI 2007), p 14401445
[62] Pradhan S, Ramshaw L, Weischedel R, MacBride J, Micciulla L (2007) Unrestricted coreference: Identifying entities and events in ontonotes. In: Proceedings of the IEEE International Conference on Semantic Computing (ICSC)
[63] Pradhan S, Ramshaw L, Marcus M, Palmer M, Weischedel R, Xue N (2011) CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Portland, OR, pp 1–27
[64] Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y (2012) CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Joint Conference on EMNLP and CoNLL - Shared Task, Jeju Island, Korea, pp 1–40
[65] Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric to semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19(1):17–30

[66] Rahman A, Ng V (2009) Supervised models for coreference resolution. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP09), Singapore, pp 968–977
[67] Rahman A, Ng V (2011) Coreference resolution with world knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011), Portland, OR, pp 814–824
[68] Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: Proceedings of the Conference on Computational Natural Language Learning (CoNLL), Boulder, CO, pp 147–155
[69] Ratinov L, Roth D (2012) Learning-based multi-sieve coreference resolution with knowledge. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012), Jeju Island, Korea, pp 1234–1244
[70] Ratinov L, Downey D, Anderson M, Roth D (2011) Local and global algorithms for disambiguation to Wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011), Portland, OR, pp 1375–1384
[71] Ravichandran D, Pantel P, Hovy E (2005) Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Ann Arbor, MI, pp 622–629
[72] Recasens M, Màrquez L, Sapena E, Martí MA, Taulé M, Hoste V, Poesio M, Versley Y (2010) Semeval task 1: Coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010), Los Angeles, CA, pp 1–8
[73] Recasens M, Can M, Jurafsky D (2013) Same referent, different words: Unsupervised mining of opaque coreferent mentions. In: Proceedings of NAACL-HLT 2013
[74] Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain, pp 1089–1090
[75] Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4):521–544
[76] Stoyanov V, Gilbert N, Cardie C, Riloff E (2009) Conundrums in noun phrase coreference resolution: Making sense of the state of the art. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL/IJCNLP 2009), Singapore, pp 656–664
[77] Stoyanov V, Cardie C, Gilbert N, Riloff E, Buttler D, Hysom D (2010) Coreference resolution with Reconcile. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp 156–161

[78] Suchanek FM, Kasneci G, Weikum G (2007) YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In: Proceedings of the 16th World Wide Web Conference (WWW 2007), Banff, Canada, pp 697–706

[79] Telljohann H, Hinrichs EW, Kübler S, Zinsmeister H, Beck K (2009) Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Tech. rep., Seminar für Sprachwissenschaft, Universität Tübingen

[80] Uryupina O, Poesio M, Giuliano C, Tymoshenko K (2011) Disambiguation and filtering methods in using Web knowledge for coreference resolution. In: Proceedings of the Twenty-Fourth International FLAIRS Conference (FLAIRS 2011)

[81] Versley Y (2006) A constraint-based approach to noun phrase coreference resolution in German newspaper text. In: Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006), Konstanz, pp 143–150

[82] Versley Y (2007) Antecedent selection techniques for high-recall coreference resolution. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prag, Tschechische Republik, pp 496–505

[83] Versley Y, Ponzetto S, Poesio M, Eidelman V, Jern A, Smith J, Yang X, Moschitti A (2008) BART: A modular toolkit for coreference resolution. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session (ACL 2008 Demo), Columbus, OH, pp 9–12

[84] Vieira R, Poesio M (2000) An empirically based system for processing definite descriptions. Computational Linguistics 26(4):539–593

[85] Wu Z, Palmer M (1994) Verb semantics and lexical selections. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994), Las Cruces, NM, pp 133–138

[86] Yang X, Su J (2007) Coreference resolution using semantic relatedness information from automatically discovered patterns. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, pp 528–535

[87] Zheng J, Vilnis L, Singh S, Choi J, McCallum A (2013) Dynamic knowledge-base alignment for coreference resolution. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL 2013), Sofia, Bulgaria, pp 153–162