Gender and Animacy Knowledge Discovery from Web-Scale N-Grams for Unsupervised Person Mention Detection[*]

Heng Ji[a] and Dekang Lin[b]

[a]Computer Science Department, Queens College and Graduate Center, City University of New York
65-30 Kissena Boulevard, Flushing, NY 11367, USA
hengji@cs.qc.cuny.edu
[b]Google, Inc.
1600 Amphitheater Parkway, Mountain View, CA 94043, USA
lindek@google.com

Abstract. In this paper we present a simple approach to discover gender and animacy knowledge for person mention detection. We learn noun-gender and noun-animacy pair counts from web-scale n-grams using specific lexical patterns, and then apply confidence estimation metrics to filter noise. The selected informative pairs are then used to detect person mentions from raw texts in an unsupervised learning framework. Experiments showed that this approach can achieve high performance comparable to state-of-the-art supervised learning methods which require manually annotated corpora and gazetteers.

Keywords: Knowledge Discovery, Mention Detection, N-Grams, Gender, Animacy

1 Introduction

The task of detecting entity mentions (references to entities) is very important to the downstream processing of information extraction such as coreference resolution and event extraction. Entity mentions can be divided into name mentions (e.g. "*John Smith*"), nominal mentions (e.g. "*president*") and pronouns (e.g. "*he*", "*she*"). Typical mention detection systems are based on supervised learning (Boschee *et al*., 2005; Zitouni and Florian, 2008) or semi-supervised learning (Ji and Grishman, 2006). Achieving really high performance for mention detection requires deep semantic knowledge and large costly hand-labeled data. Many systems also exploited lexical gazetteers such as census data with gender information. However, such knowledge is relatively static (it is not updated during the extraction process), expensive to construct, and doesn't include any probabilistic information.

Mention detection is by definition a semantic task: for example, a phrase is a person mention if it refers to a real-world person entity. We should thus expect a successful mention detection system to exploit world knowledge, in order to resolve hard cases. For example, if a reflexive pronoun (e.g. "himself") is bound by a phrase in its governing category (Haegeman, 1994), then this phrase is likely to be a person mention (masculine or feminine). In addition, a person mention usually has a life and therefore is likely to be animate (Cobuild, 1995). Therefore, if we could automatically discover a large knowledge base of *gender* and *animacy* properties for all possible noun phrases, it will be a valuable resource for person mention detection.

[*] This work was supported by National Science Foundation and Google, Inc. through the Johns Hopkins University 2009 Summer Workshop on Unsupervised Acquisition of Lexical Knowledge from N-Grams. The first author's research was also sponsored by CUNY Research Enhancement Program and GRTI Program, and the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

In this paper we will glean these two powerful lexical properties – gender and animacy – for person mention detection. Further progress will likely be aided by flexible frameworks for representing and using the information provided by this kind of properties. We shall discover these properties from web-scale Google n-gram data and use them to detect person mentions in an unsupervised learning fashion. Such methods allow us to compensate for the absence of annotated training data and semantic resources. The derived properties may include a lot of noise, and thus we will introduce several confidence estimation methods and experiment with various patterns for knowledge discovery. The contributions of this paper are two-fold: (1) the first attempt to discover gender and animacy knowledge from web-scale n-grams; (2) the first work on detecting entity mentions based on unsupervised knowledge discovery.

The rest of this paper is structured as follows. Section 2 describes our main research task and experimental setting. Section 3 motivates our approach based with error analysis of traditional supervised learning. Section 4 and Section 5 then present the detailed knowledge discovery process from n-grams and using them for mention detection. Section 6 presents experimental results. Section 7 briefly reviews the previous research on the discovery and the use of gender and animacy knowledge. Section 8 then concludes the paper and sketches our future work.

2 Terminology and Task Definition

The mention detection task we are addressing is that of the Automatic Content Extraction (ACE) evaluations (NIST, 2005). ACE defines the following terminology:

entity: an object or a set of objects in one of the semantic categories of interest: person, location, geo-political, organization, facility, vehicle and weapon

person name mention: a reference by name to a person entity

person nominal mention: a reference by a common noun or noun phrase to a person entity

For example, for a sentence: "*John Smith is a famous screenwriter in LA.*", a mention detector should identify "*John Smith*" as a person name mention and "*[a famous] screenwriter*" as a nominal mention with "*screenwriter*" as head. In this paper we consider a mention as correct only if its type and head exactly match the reference.

3 Error Analysis of Supervised Learning Methods for Mention Detection

We begin our error analysis with an investigation of a state-of-the-art English mention detection system based on supervised learning (Grishman *et al*., 2005), decomposing the errors into name mention and nominal mention detection errors.

The baseline name tagger is based on Hidden Markov Model (HMM) trained on about 1375 ACE documents. The HMM includes six states for each of the seven entity types defined in ACE, as well as a not-a-name state. These six states correspond to the token preceding the name; the single name token (for names with only one token); the first token of the name; an internal token of the name (neither first nor last); the last token of the name; and the token following the name. To detect nominal mentions, the system starts from a HMM based part-of-speech tagger and a Maximum Entropy (MaxEnt) based noun phrase chunker trained from the Penn Treebank. The main features used in chunking are the bigram conjunctions of POS features. Then the person nominal mentions are detected by matching the noun phrase heads against a list of 121 title words and 29,425 person nominal mentions from ACE training corpora. In addition, the system exploited a manually constructed name gazetteer including 245,615 names and the census data including 5,014 person-gender pairs.

Some mentions can be correctly identified using the above supervised learning methods, but these methods still suffer from the limited availability of large annotated corpora and semantic resources and therefore leave a large number of mentions unidentified. The F-measure of name mention detection is about 84.5% and nominal mention is only about 75%. For example, in the following sentence in Figure 1, we can see that the name mention error types are quite diverse -

the system mistakenly classified two geo-political names *“Faisalabad”* and *“Sahiwal”* as persons, and tagged a spurious name mention *“Catholic Bishop”* because of its capitalization feature, and missed a rare person name *“Ayub Masih”*. For nominal mentions, there are more missing errors than other error types because a lot of them rarely appear in the training data, such as “*supremo*”, “*shepherd*”, “*prophet*”, “*sheikh*”, “*Imam*”, “*overseer*”, “*oligarchs*” and “*Sheikh*”. However, assuming we have an extremely large unlabeled corpus, such as all the data on the web, most of these instances must have occurred. Therefore the remaining question is – can we automatically discover these mentions from very large data by effective semantic constraints while not introduce too much noise? We shall describe the approaches to discover gender and animacy properties (Section 4) and incorporate them into unsupervised learning (Section 5) respectively.

> *Reference: Faisalabad's Catholic Bishop* **<PER>John Joseph</PER>***, who had been campaigning against the law, shot himself in the head outside a court in Sahiwal district when the judge convicted Christian* **<PER>Ayub Masih</PER>** *under the law in 1998.*
>
> *System:* **<PER>Faisalabad</PER>***'s* **<PER>Catholic Bishop</PER> <PER>John Joseph</PER>***, who had been campaigning against the law, shot himself in the head outside a court in* **<PER>Sahiwal</PER>** *district when the judge convicted Christian Ayub Masih under the law in 1998.*

Figure 1: Name Mention Detection Error Examples from Supervised Learning

4 Gender and Animacy Knowledge Discovery from Web-scale N-Grams

Since the gender and animacy properties of words are highly correlated with whether a noun phrase is a person mention, these properties are expected to be very useful for identifying person mentions. In this paper we take use of Google n-gram (n=5) corpus Version II, which can be viewed as a compressed summary of the web, to discover such properties in an offline fashion. Google n-gram Version II includes 207 billion tokens selected from the LDC-released Version I, consisted of 1.2 billion 5-grams extracted from about 9.7 billion sentences. All these 5-grams are automatically annotated with part-of-speech (POS) tags based on their original sentences.

Table 1: Patterns to Discover Gender and Animacy Properties

Property	Name	*target [#]*	context	*Pronoun*	Example
Gender	Conjunction-Possessive	noun[292,212] | capitalized [162,426]	conjunction	his|her|its|their	*writer* and *his*
	Nominative-Predicate	noun [53,587]	am|is|are| was|were|be	he|she|it|they	*he* is a *writer*
	Verb-Nominative	noun [116,607]	verb	he|she|it|they	*writer* thought *he*
	Verb-Possessive	noun [88,577]| capitalized [52,036]	verb	his|her|its|their	*writer* bought *his*
	Verb-Reflexive	noun [18,725]	verb	himself|herself| itself|themselves	*writer* explained *himself*
Animacy	Relative-Pronoun	(noun|adjective) & not after (preposition| noun|adjective) [664,673]	comma| empty	who|which| where|when	*writer*, *who*

We design the patterns in Table 1 to extract gender and animacy frequencies for each pair of *target-pronoun* from Google 5-grams. Most of the gender patterns follow the general idea in (Bergsma, 2005).

For example, in the “Conjunction-Possessive” pattern, we count the possessive pronouns following a conjunction word after the nouns in order to get their gender properties (e.g. if “writer and his” appears frequently then it indicates that “writer” is a often a male); and in the “Relative-pronoun” pattern we count the relative pronouns after nouns to determine their animacy properties (e.g. if “writer, who” appears frequently then it indicates that “writer” is often animate). When the target word is capitalized we use these properties to detect name mentions, otherwise to detect nominal mentions. In the target column we also present the number of discovered targets by each pattern. In total we discovered 784,170 targets with gender property and 664,673 targets with animacy property. These semantic resources are freely available for research purposes: http://nlp.cs.qc.cuny.edu/ngram_genderanimacy.zip.

We then map the discovered *target-pronoun* pairs into corresponding properties in Table 2. The basic intuition of our method is that if a target indicates masculine/feminine/animate with high confidence, then it’s likely to be a person mention.

Table 2: Lexical Property Mapping

Property	Pronoun	Value
Gender	his|he|himself	masculine
	her|she|herself	feminine
	its|it|itself	neutral
	their|they|themselves	plural
Animacy	who	animate
	which|where|when	non-animate

Table 3: Gender Property Examples

Target	masculine	feminine	neutral	Plural
John Joseph	32	0	0	0
Haifa	21	19	92	15
screenwriter	144	27	0	0
Fish	22	41	1741	1186

Table 4: Animacy Property Examples

target	Animate	Non-Animate		
	Who	when	where	which
Supremo	24	0	0	0
shepherd	807	24	0	56
Prophet	7372	1066	63	1141
Imam	910	76	0	57
oligarchs	299	13	0	28
Sheikh	338	11	0	0

Table 3 presents some examples with their gender frequencies. We can clearly see that the person mentions such as "*John Joseph*" and "*screenwriter*" only have "*masculine/feminine*" properties; while "*Haifa*" appears mostly as neutral and "*fish*" appears as neutral/plural, which indicate that they are unlikely to be person mentions. Table 4 shows the animacy statistics for some of the nominal mentions missed by the baseline supervised learning model. We can see that all of them appear as animate much more frequently than inanimate in n-grams, and thus this property can also be used to identify person mentions effectively.

5 Using Gender and Animacy Properties in Unsupervised Learning

Most of the prior work of using knowledge sources focused on encoding them as additional features in supervised learning models. However, for some domains such as financial analysis very few annotated training corpora are available for mention detection. Therefore in this paper we are more interested in investigating how much we can achieve on this task by only using the semantic knowledge discovered from Google n-grams, namely in a completely unsupervised learning framework. We shall present the overall procedure in section 5.1 and then focus on discussing the possible confidence estimation metrics in section 5.2

5.1 Overall Procedure

Figure 2 depicts the general procedure of our approach.

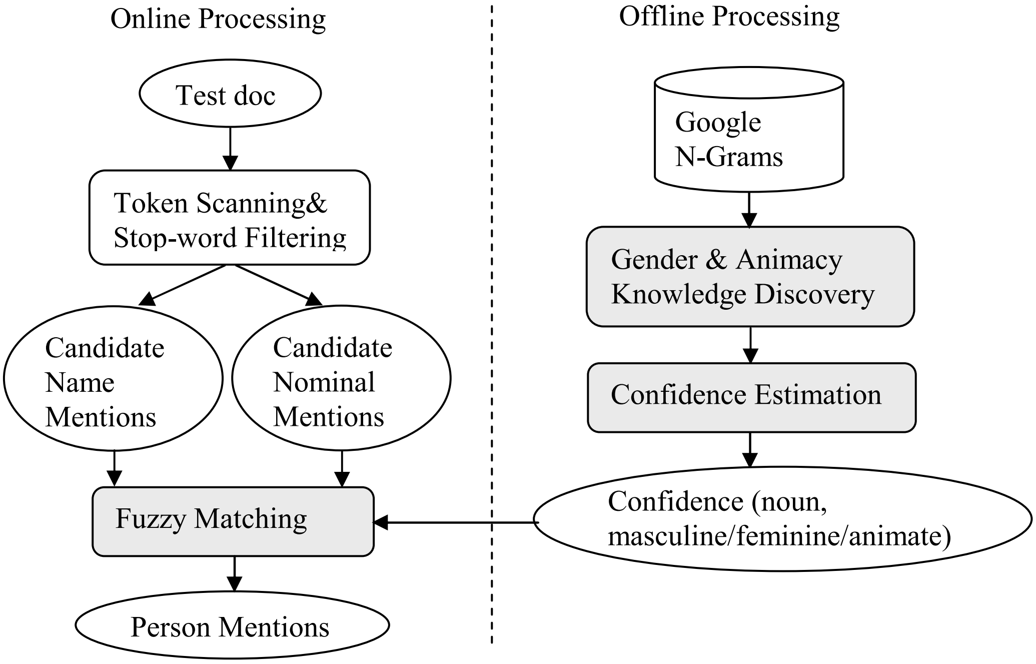
Online Processing
Offline Processing
Test doc
Token Scanning& Stop-word Filtering
Candidate Name Mentions
Candidate Nominal Mentions
Fuzzy Matching
Person Mentions
Google N-Grams
Gender & Animacy Knowledge Discovery
Confidence Estimation
Confidence (noun, masculine/feminine/animate)

Figure 2: Overall Procedure of Unsupervised Learning for Person Mention Detection

The input text is segmented into sentences and scanned against stop word lists to generate candidate mentions. Each string of three or fewer non-stop tokens is considered as a candidate mention; if all the tokens in the string are capitalized then it's treated as a name mention candidate and otherwise as a nominal mention candidate. In addition, for name mention detection we further filter dates, numbers and title words from candidates.

For each candidate mention string $str\ [token_1 ... token_n]$, we look it up in the gender and animacy knowledge base discovered from Google n-grams. If it matches one of the following conditions, it's generated as a person mention:

- **Full matching**
  Confidence (*str*, *masculine/feminine/animate*) > δ
- **Composite matching**
  For any *i* in *[1, n]*, Confidence ($token_i$, *masculine/feminine/animate*) > δ
- **Relaxed matching**
  For any *i* and *j* in *[1, n]*, Confidence ($token_i$, *masculine/feminine/animate*) > δ and Confidence ($token_j$, *masculine/feminine/animate*) > δ

The following Table 5 lists some examples for each of the above matching methods. For instance, although "*Qawasmi*" doesn't exist in the knowledge base, we can still identify "*Mahmoud Salim Qawasmi*" as a name mention because both "*Mahmoud*" and "*Salim*" have the properties of "*masculine/feminine*" with high confidence values.

Table 5: Property Matching Examples

Mention candidate	Matching Method	String for matching	Property Frequency			
			masculine	feminine	neutral	plural
John Joseph	Full Matching	John Joseph	**32**	**0**	0	0
Ayub Masih	Composite Matching	Ayub	**87**	**0**	0	0
		Masih	**117**	**0**	0	0
Mahmoud Salim Qawasmi	Relaxed Matching	Mahmoud	**159**	**13**	0	0
		Salim	**188**	**13**	0	0
		Qawasmi	0	0	0	0

5.2 Confidence Estimation

There was a time when lack of data was a problem in many tasks. In our approach, the contrary is true – the extremely large n-grams provide us high coverage of candidate mentions but at the same time bring a lot of noise. Therefore we need to explore various effective confidence estimation metrics in order to separate the "wheat" from the "chaff". We rank the properties for each noun according to their frequencies (from high to low): $[f_1 ... f_k]$, and attempt the following different metrics.

- $percentage = \frac{f_1}{\sum_{i=1}^{k} f_i}$

As the simplest and most intuitive metric, percentage reflects the confidence of a property among the overall ranked list.

- $margin = \frac{f_1 - f_2}{f_2}$

The second alternative we consider is the "margin" metric which is widely used in the active learning community (e.g. Jones *et al.*, 2003; Riccardi *et al.*, 2004), measuring the difference between the best property and the second best property. If the margin is larger, then the best property is more likely to be correct.

- *margin&frequency* $= \frac{f_1}{f_2} \times \log(f_1)$

In some cases we may want to add some weights to those frequent nouns, therefore we propose the third metric by adding frequency information to margin. This is similar to the relevancy snippet selection metric described in (Riloff, 1996).

6 Experimental Results

In this section we present the results of applying gender and animacy properties to detect person mentions.

6.1 Data

We use 10 newswire texts from ACE 2005 training corpora as our development set, and then conduct blind test on a separate set of 50 ACE 2005 newswire texts. The test set includes 555 person name mentions and 900 person nominal mentions.

6.2 Impact of Confidence Metrics

Since each pattern involves confidence estimation metrics, it's important to select effective thresholds. As an example, we select the thresholds (δ_k with k=1~3) for various confidence metrics by optimizing the F-measure score of the Conjunction-Possessive pattern on the development set, as shown in Figure 3. Each curve in Figure 3 shows the effect on name mention detection precision and recall of varying the threshold for each confidence metric.

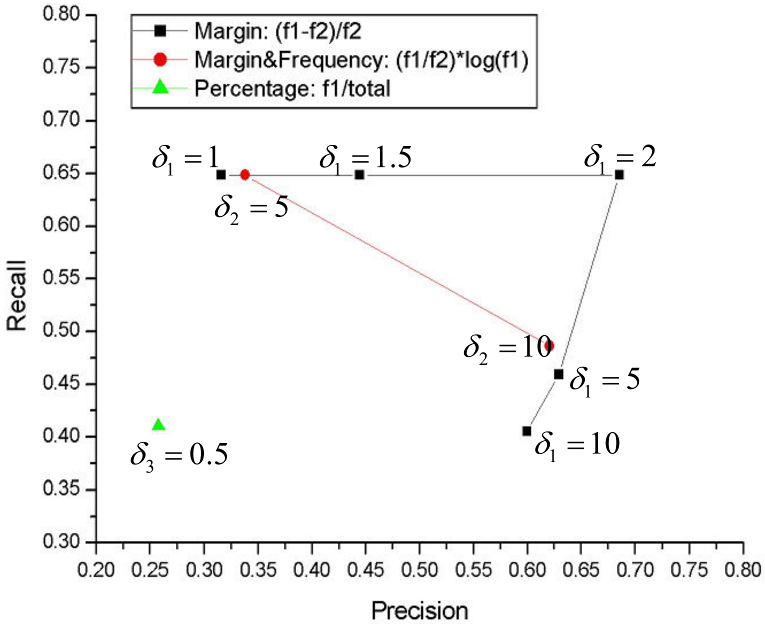

Figure 3: Optimizing Confidence Metrics of Conjunction-Possessive Pattern for Gender Discovery based Name Mention Detection in Development Set

We can see that the best F-measure can be obtained on the development set by setting the threshold $\delta_1 = 2$ for the margin metric. After we optimize these thresholds on the development set, we use them directly for blind testing to report the final experimental results in section 6.4.

We found that other metrics were less reliable than margin mainly because the over-weighting of frequency information caused more spurious errors. For example, for the word "*under*", the results of using the margin&frequency metric are as follows: *margin&frequency(under, masculine)* = 30, *margin&frequency(under, feminine)* =233, *margin&frequency(under, neutral)* = 15, *margin&frequency(under, plural)* = 49, and so "*under*" will be mistakenly identified as a person mention. We believe further improvement can be achieved if we take into account the global frequency information of candidates in the overall n-grams without pattern restrictions.

6.3 Impact of Knowledge Sources

We investigate the contribution of each individual pattern separately on mention detection. Table 6 below presents the performance of nominal mention detection. The results indicate that the properties discovered by any single pattern cannot yield satisfying performance, but consistent improvements were achieved as we add the diverse patterns gradually. Among these patterns the Nominative-Predicate and Verb-Possessive patterns for gender discovery and the Relative-Pronoun pattern for animacy discovery had the largest impact, improving recall significantly.

Table 6: Impact of Diverse Patterns for Nominal Mention Detection on Development Set

Patterns		Nominal Mention Detection		
		Precision (%)	Recall (%)	F-Measure (%)
Gender	Conjunction-Possessive	78.57	10.28	18.18
	+Nominative-Predicate	78.57	20.56	32.59
	+Verb-Nominative	65.85	25.23	36.49
	+Verb-Possessive	55.71	36.45	44.07
	+Verb-Reflexive	64.41	35.51	45.78
Animacy	+Relative-Pronoun	63.33	71.03	66.96

6.4 Overall Performance

Table 7 shows the overall Precision, Recall and F-Measure scores on the blind test set, using the baseline supervised learning method as described in section 3 and our new unsupervised learning method based on knowledge discovery.

Table 7: Overall Performance of Person Mention Detection on Test Set

Task	Method	Precision (%)	Recall (%)	F-Measure (%)
Name Mention Detection	Supervised Learning	88.24	81.08	84.51
	Unsupervised Learning Using Knowledge Discovery from Web-scale N-Grams	87.05	82.34	84.63
Nominal Mention Detection	Supervised Learning	85.93	70.56	77.49
	Unsupervised Learning Using Knowledge Discovery from Web-scale N-Grams	71.20	85.18	77.57

Table 7 indicates that our unsupervised learning method based on knowledge discovery achieved comparable performance with the traditional supervised learning model for both name and nominal mention detection. Our approach has the advantage of higher coverage on low frequency mentions. For example, it successfully identified all the nominal mentions in Table 4 which were missed by the supervised learning model. However, due to the noise produced from n-grams and the limited use of specific contexts, our method had more loss in precision. Typically some organizations named after people such as "JP Morgan" were mistakenly identified as person mentions because of their high confidence with *masculine/feminine/animate* properties. Nevertheless, given the fact that we didn't use any manually annotated training data or semantic resources, these results are promising. Furthermore, we believe a semi-supervised learning framework incorporating the discovered knowledge will further boost the performance because the difficult cases tackled by these two methods are complementary.

7 Related Work

Our method exhibits a fundamental advantage over supervised learning algorithm (including Boschee *et al*., 2005; Ji and Grishman, 2006; Zitouni and Florian, 2008) as it does not require costly hand-labeled training data. It thrives on web-scale Google n-gram data and discovers semantic knowledge corresponding to the task of mention detection.

The use of gender information stems from a lot of prior work on pronoun resolution. Most of these methods (e.g. Ge *et al.,* 1998; Cardie and Wagstaff, 1999) encoded the gender information as hard constraints. Hale and Charniak (1998) obtained gender statistics by using an anaphora algorithm on a large corpus. Bergsma *et al*. (2005, 2009a) mined gender information from the web and parsed corpora and incorporated gender probabilities as additional features in supervised learning. To the best of our knowledge, this is the first work on exploiting gender information for mention detection and in an unsupervised learning framework. Some very recent work used Google n-gram data for other NLP tasks such as lexical disambiguation (Bergsma *et al*., 2009b). Limited prior work has used manually constructed knowledge resources such as WordNet for Animacy Discovery (Evans and Orasan, 2000). Our offline strategy for acquiring gender and animacy information for online mention detection is similar to that for question answering described in Fleischman *et al*. (2003). And our approach of using pronoun context to improve mention detection is similar to the idea of refining name tagging based on coreference feedback in (Ji *et al*., 2005).

8 Conclusion

Using mention detection as a case study, we have demonstrated that unsupervised learning methods can achieve comparable performance for some particular tasks if we discover semantic knowledge corresponding to each task. Our method harnesses the probabilistic lexical properties such as gender and animacy discovered from web-scale n-grams, and therefore can identify more rare mentions than the traditional supervised learning methods based on limited and static semantic resources. Also as an unsupervised learning approach it performs surprisingly well especially on recall. We have also proved that the properties discovered from large n-grams are not in themselves sufficient: we must acquire 'clean' knowledge by effective confidence estimation and parameter tuning. In the future we are interested in exploring the same idea of knowledge discovery for other more complicated IE tasks such as event extraction. In addition we will aim to extend our approach to other languages for which Google n-grams are available, including Chinese and Japanese.

References

Bergsma, S. 2005. Automatic Acquisition of Gender Information for Anaphora Resolution. *Proc. Canadian AI 2005*.

Bergsma, S., D. Lin and R. Goebel. 2009a. Glen, Glenda or Glendale: Unsupervised and Semi-supervised Learning of English Noun Gender. *Proc. CoNLL 2009*.

Bergsma S., D. Lin and R. Goebel. 2009b. Web-Scale N-gram Models for Lexical Disambiguation. *Proc. IJCAI 2009*.

Boschee, E., R. Weischedel and A. Zamanian. 2005. Automatic Evidence Extraction. *Proc. International Conference on Intelligent Analysis*.

Cardie, C. and K. Wagstaff. 1999. Noun Phrase Coreference as Clustering. *Proc. Joint SIRDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Cobuild. 1995. *English Collocations on CD-ROM*. Harper Collins, London.

Evans, R. and C. Orasan. 2000. Improving Anaphora Resolution by Identifying Animate Entities in Texts. *Proc. the Discourse Anaphora and Reference Resolution Conference.*

Fleischman, M., E. Hovy and A. Echihabi. 2003. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. *Proc. ACL 2003*.

Ge, N., J. Hale and E. Charniak. 1998. A Statistical Approach to Anaphora Resolution. *Proc. the Sixth Workshop on Very Large Corpora*.

Grishman, R., D. Westbrook and A. Meyers. 2005. NYU's English ACE 2005 System Description. *Proc. ACE 2005 Evaluation Workshop*.

Haegeman, L. 1994. *Introduction to Government and Binding Theory (Second Edition)*. Basil Blackwell, Cambridge, UK.

Hale, J. and E. Charniak. 1998. Getting Useful Gender Statistics from English Text. *Tech Report CS-98-06*. Brown University.

Ji, H. and R. Grishman. 2006. Data Selection in Semi-supervised Learning for Name Tagging. *Proc. COLING/ACL 06 Workshop on Information Extraction Beyond Document*.

Ji, H., D. Westbrook and R. Grishman. 2005. Using Semantic Relations to Refine Coreference Decisions. *Proc. HLT/EMNLP 05*.

Jones, R., R. Ghani, T. Mitchell and E. Riloff. 2003. Active Learning for Information Extraction with Multiple View Feature Sets. *Proc. ECML-03 Workshop on Adaptive Text Extraction and Mining*.

NIST. 2005. Automatic Content Extraction. *http://www.nist.gov/speech/tests/ace/*

Riccardi, G., D. Hakkani-Tür, G. Tur, Adaptive Learning: From Supervised to Active Learning of Statistical Models for Natural Language and Speech Processing. *Proc. ACL 2004*.

Riloff, E. 1996. Automatically Generating Extraction Patterns from Untagged Text. *Proc. AAAI 1996.*

Zitouni, I. and R. Florian. 2008. Mention Detection Crossing the Language Barrier. *Proc. EMNLP 2008.*