Challenges from Information Extraction to Information Fusion

Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
hengji@cs.qc.cuny.edu

Abstract

Information Extraction (IE) technology is facing new challenges of dealing with large-scale heterogeneous data sources from different documents, languages and modalities. Information fusion, a new emerging area derived from IE, aims to address these challenges. We specify the requirements and possible solutions to perform information fusion. The issues include redundancy removal, contradiction resolution and uncertainty reduction. We believe this is a critical step to advance IE to a higher level of performance and portability.

1 Introduction

Latest development of Information Extraction (IE) techniques has made it possible to extract 'facts' (entities, relations and events) from unstructured documents, and converting them into structured representations (e.g. databases). Once the collection grows beyond a certain size, an issue of critical importance is how a user can monitor a compact knowledge base or identify the interesting portions without having to (re) read large amounts of facts. In this situation users are often more concerned with the speed in which they obtain results, rather than obtaining the exact answers to their queries (Jagadish et al., 1999). The facts extracted from heterogeneous data sources (e.g. text, images, speech and videos) must then be integrated in a knowledge base, so that it can be queried in a uniform way. This provides unparalleled challenges and opportunities for improved decision making.

Data can be noisy, incorrect, or misleading. Unstructured data, mostly text, is difficult to interpret. In practice it is often the case that there are multiple sources which need to be extracted and compressed. In a large, diverse, and interconnected system, it is difficult to assure accuracy or even coherence among the data sources. In this environment, traditional IE would be of little value. Most current IE systems focus on processing a single document and language, and are customized for a single data modality. In addition, automatic IE systems are far from perfect and tend to produce errors.

Achieving really advances in IE requires that we take a broader view, one that looks outside a single source. We feel the time is now ripe to incorporate some information integration techniques in the database community (e.g. Seligman et al., 2010) to extend the IE paradigm to real-time information fusion and raise IE to a higher level of performance and portability. This requires us to work on a more challenging problem of *information fusion* - to remove redundancy, resolve contradictions and uncertainties by multiple information providers and design a general framework for the veracity analysis problem. The goal of this paper is to lay out the current status and potential challenges of information fusion, and suggest the following possible research avenues.

- **Cross-document:** We will discuss how to effectively aggregate facts across documents via entity and event coreference resolution.
- **Cross-lingual:** A shrinking fraction of the world's Web pages are written in English, and so the ability to access pages across a range of languages is becoming increasingly important for many applications. This need can be addressed in part by cross-lingual information fusion. We will discuss the chal-

Coling 2008: Poster Volume, pages 507–515,
Beijing, August 2010

lenges of extraction and translation respectively.

- **Cross-media:** Advances in speech and image processing make the application of IE possible on other data modalities, beyond traditional textual documents.

2 Cross-Document Information Fusion

Most current IE systems focus on processing one document at a time, and except for coreference resolution, operate one sentence at a time. The systems make only limited use of 'facts' already extracted in the current document. The output contains rich structures about entities, relations and events involving such entities. However, due to noise, uncertainty, volatility and unavailability of IE components, the collected facts may be incomplete, noisy and erroneous. Several recent studies have stressed the benefits of using information fusion across documents. These methods investigate quite different angles while follow a common research theme, namely to exploit global background knowledge.

2.1 Information Inference

Achieving really high performance (especially, recall) of IE requires deep semantic knowledge and large costly hand-labeled data. Many systems also exploited lexical gazetteers. However, such knowledge is relatively static (it is not updated during the extraction process), expensive to construct, and doesn't include any probabilistic information. Error analysis on relation extraction shows that a majority (about 78%) of errors occur on nominal mentions, and more than 90% missing errors occur due to the lack of enough patterns to capture the context between two entity mentions. For instance, to describe the "located" relation between a bomber and a bus, there are more than 50 different intervening strings (e.g. "killed many people on a", "'s attack on a", "blew apart a", "blew himself up on a", "drove his explosives-laden car into a", "had rigged the", "set off a bomb on a", etc.), but the ACE[1] training corpora only cover about 1/3 of these expressions.

Several recent studies have stressed the benefits of using information redundancy on estimating the correctness of the IE output (Downey et al., 2005), improving disease event extraction (Yangarber, 2006), Message Understanding Conference event extraction (Mann, 2007; Patwardhan and Riloff, 2009) and ACE event extraction (Ji and Grishman, 2008). This approach is based on the premise that many facts will be reported multiple times from different sources in different forms. This may occur both within the same document and within a cluster of topically related and successive documents. Therefore, by aggregating similar facts across documents and conducting statistical global inference by favoring interpretation consistency, enhanced extraction performance can be achieved with heterogeneous data than uniform data.

The underlying hypothesis of cross-document inference is that the salience of a fact should be calculated by taking into consideration both its confidence and the confidence of other facts connected to it, which is inspired by PageRank (Page et al., 1998) and LexRank (Erkan and Radev, 2004). For example, a vote by linked entities which are highly voted on by other entities is more valuable than a vote from unlinked entities. There are two major heuristics: (1) *an assertion that several information providers agree on is usually more trustable than that only one provider suggests*; and (2) *an information provider is* trustworthy *if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy providers*. (Yin et al., 2008) used the above heuristics in a progressive, iterative enhancement process for information fusion.

The results from the previous work are promising, but the heuristic inferences are highly dependent on the order of applying rules, and the performance may have been limited by the thresholds which may overfit a small development corpus. One promising method might be using Markov Logic Networks (Richardson and Domingos, 2006), a statistical relational learning language, to model these global inference rules more declaratively. Markov Logic will make it possible to compactly specify probability distributions over the complex relational inferences. It can capture non-deterministic (soft) rules that tend to hold among facts but do not have to. Exploiting this approach will also provide greater flexibility to incorporate additional linguistic and world knowledge into inference.

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

The information fused across documents can be represented as an information network (Ji, 2009) in which entities can be viewed as vertices on the graph and they can be connected by some type of static relationship (e.g. those attributes defined in NIST TAC-KBP task (McNamee and Dang, 2009)), or as a temporal chain linking dynamic events (e.g. Bethard and Martin, 2008; Chambers and Jurafsky, 2009; Ji et al., 2009a). The latter representation is more attractive because business or international affairs analysts often review many news reports to track people, companies, and government activities and trends. The query logs from the commercial search engines show that there is a fair number of news related queries (Mishne & de Rijke, 2006), suggesting that blog search users have an interest in the blogosphere response to news stories as they develop. For example, (Ji et al., 2009a) extracted centroid entities and then linked events centered around the same centroid entities on a time line.

Temporal ordering is a challenging task in particular because about half of the event mentions don't include explicit time arguments. The text order by itself is a poor predictor of chronological order (only 3% temporal correlation with the true order). Single-document IE technique can identify and normalize event time arguments from the texts, which results in a much better correlation score of 44% (Ji et al., 2009a). But this is still far from the ideal performance for real applications. In order to alleviate this bottleneck, a possible solution is to exploit global knowledge from the related documents and Wikipedia, and related events to recover and predict some implicit time arguments (Filatova and Hovy, 2001; Mani et al., 2003; Mann, 2007; Eidelman, 2008; Gupta and Ji, 2009).

2.2 Coreference Resolution

One of the key challenges for information fusion is cross-document entity coreference – precise clustering of mentions into correct entities. There are two principal challenges: the same entity can be referred to by more than one name string and the same name string can refer to more than one entity. The recent research has been mainly promoted in the web people search task (Artiles et al., 2007) such as (Balog et al., 2008), ACE2008 such as (Baron and Freedman, 2008) and NIST TAC KBP (McNamee and Dang, 2009) evaluations. Interestingly, the quality of information can often be improved by the fused fact network itself, which can be called as self-boosting of information fusion. For example, if two GPE entities are involved in a "conflict-attack" event, then they are unlikely to be connected by a "part-whole" relation; "Mahmoud Abbas" and "Abu Mazen" are likely to be coreferential if they get involved in the same "life-born" event. Some prior work (Ji et al., 2005; Jing et al., 2007) demonstrated the effectiveness of using semantic relations to improve entity coreference resolution; while (Downey et al., 2005; Sutton and McCallum, 2004; Finkel et al., 2005; Mann, 2007) experimented with information fusion of relations across multiple documents. The TextRunner system (Banko et al., 2007) can collapse and compress redundant facts extracted from multiple documents based on coreference resolution (Yates and Etzioni, 2009), semantic similarity computation and normalization.

Two relations are central for event fusion: *contradiction* – part of one event mention contradicts part of another, and *redundancy* – part of one event mention conveys the same content as (or is entailed by) part of another. Once these central relations are identified they will provide a basis for identifying more complex relations such as elaboration, presupposition or consequence. It is important to note that redundancy and contradiction among event mentions are *logical* relations that are not captured by traditional topic-based techniques for similarity detection (e.g. Brants and Stolle, 2002). Contradictions also arise from complex differences in the structure of assertions, discrepancies based on world-knowledge, and lexical contrasts. Ritter et al. (2009) described a contradiction detection method based on functional relations and pointed out that many contradictory fact pairs from the Web appear consistent, and that requires background knowledge to predict.

Assessing event coreference is essential: for texts to contradict, they must refer to the same event. Event coreference resolution is more challenging than entity coreference because each linking decision needs to be made based upon the overall similarity of the event trigger and multiple arguments. Hasler and Orasan (2009)

further found that in many cases even coreferential even arguments are not good indicators for event coreference.

Earlier work on event coreference resolution (e.g. Bagga and Baldwin, 1999) was limited to several MUC scenarios. Recent work (Chen et al., 2009) focus on much wider coverage of event types defined in ACE. The methods from the knowledge fusion community (e.g. Appriou et al., 2001; Gregoire, 2006) mostly focus on resolving conflicts rather than identifying them (i.e. inconsistency problem rather than ambiguity). These approaches allow the conflicts to be resolved in a straightforward way but they rely on the availability of meta-data (e.g., distribution of weights between attributes, probability assignment etc.). However, it is not always clear where to get this meta-data.

The event attributes such as Modality, Polarity, Genericity and Tense (Sauri et al., 2006) will play an important role in event coreference resolution because two event mentions cannot be coreferential if any of the attributes conflict with each other. Such attempts have been largely neglected in the prior research due to the low weights of attribute labeling in the ACE scoring metric. (Chen et al., 2009) demonstrated that simple automatic event attribute labeling can significantly improve event coreference resolution. In addition, some very recent work including (Nicolae and Nicolae, 2006; Ng, 2009; Chen et al., 2009) found that graph-cut based clustering can improve coreference resolution. The challenge lies in computing the affinity matrix.

3 Cross-Lingual Information Fusion

Cross-lingual comparable corpora are also prevalent now because almost all the influential events can be reported in multi-languages at the first time, but probably in different aspects. Therefore, linked fact networks can be constructed and lots of research tasks can benefit from such structures. Since the two networks are similar in structure but not homogeneous, we can do alignment and translation which may advance information fusion. Cross-lingual information fusion is concerned with technologies that fuse the information available in various languages and present the fused information in the user-preferred language. The following fundamental cross-lingual IE pipelines can be employed: (1) Translate source language texts into target language, and then run target language IE on the translated texts. (2) Run source language IE on the source language texts, and then use machine translation (MT) word alignments to translate (project) extracted information into target languages. Regardless of the different architectures, both pipelines are facing the following challenges from extraction and translation.

3.1 Extraction Challenges

Some recent fusion work focus on cross-lingual interaction and inference to improve both sides synchronously, beyond the parallel comparisons of cross-lingual IE pipelines in (e.g. Riloff et al., 2002). One of such examples is on cross-lingual co-training (e.g. Cao et al., 2003; Chen and Ji, 2009). In co-training (Blum and Mitchell, 1998), the uncertainty of a classifier is defined as the portion of instances on which it cannot make classification decisions. Exchanging tagged data in bootstrapping can help reduce the uncertainties of classifiers. The cross-lingual fusion process satisfies the co-training algorithm's assumptions about two views (in this case, two languages): (1) the two views are individually sufficient for classification (IE systems in both languages were learned from annotated corpora which are enough for reasonable extraction performance); (2) the two views are conditionally independent given the class (IE systems in different languages may use different features and resources).

(Cao et al., 2003) indicated that uncertainty reduction is an important factor for enhancing the performance of co-training. It's important to design new uncertainty measures for representing the degree of uncertainty correlation of the two classifiers in co-training. (Chen and Ji, 2009) proposed a new co-training framework using cross-lingual information projection. They demonstrated that this framework is particularly effective for a challenging IE task which is situated at the end of a pipeline and thus suffers from the errors propagated from upstream processing and has low-performance baseline.

3.2 Translation Challenges

Because the facts are aggregated from multiple languages, the translation errors will bring us great challenges. However, in order to extend

cross-lingual information fusion techniques to more language pairs, we can start from the much more scalable task of "information" translation (Etzioni et al., 2007). The additional processing may take the form of machine translation (MT) of extracted facts such as names and events. IE tasks performed notably worse on machine translated texts than on texts originally written in English, and error analysis indicated that a major cause was the low quality of name translation (Ji et al., 2009b). Traditional MT systems focus on the overall fluency and accuracy of the translation but fall short in their ability to translate certain informationally critical words. In particular, it appears that better entity name translation can substantially improve cross-lingual information fusion.

Some recent work (e.g. Klementiev and Roth, 2006; Ji, 2009) has exploited comparable corpora to enhance information translation. There are no document-level or sentence-level alignments across languages, but important facts such as names, relations and events in one language in such corpora tend to co-occur with their counterparts in the other. (Ji, 2009) used a bootstrapping approach to align the information networks from bilingual comparable corpora, and discover name translations and extract relations links simultaneously. The general idea is to start from a small seed set of common name pairs, and then rely on the link attributes to align their related names. Then the new name translations are added to the seed set for the next iteration. This bootstrapping procedure is repeated until no new translations are produced. This approach is based on graph traverses and doesn't need a name transliteration module to serve as baseline, or compute document-wise temporal distributions.

The novelty of using comparable corpora lies in constructing and mining multi-lingual information fusion framework which is capable of self-boosting. First, this approach can generate information translation pairs with high accuracy by using a small seed set. Second, the shortcomings of traditional approaches are due to their limited use of IE techniques, and this approach can effectively integrate extraction and translation based on reliable confidence estimation. Third, compared to bitexts this approach can take advantage of much less expensive comparable corpora. This approach can be extended to foster the research in other aspects for information fusion. For example, the aligned sub-graphs with names, relations and events can be used to reduce information redundancy; the outlier (misaligned) sub-graphs can be used to detect the novel or local information described in one language but not in the other after the fusion process. It does happen that the two persons have been explicitly reported as Father and Son relationship in one language, but in the other language, they are just reported as two common persons.

4 Cross-Media Information Fusion

The research challenges discussed so far concerned with textual data. Besides written texts, ever-increasing human generated data is available as speech recordings, microblogs, images and videos. We now discuss how to develop techniques for fusing a variety of media sources. State-of-the-art IE techniques have been developed primarily on newspaper articles and a few web texts, and it is not clear how systems would perform on other sources and how to integrate all available information.

4.1 Coreference Resolution

The main challenge is on designing a coherent information fusion framework that is able to exploit information across different parts of multimedia documents and link them via cross-media coreference resolution. The framework will handle multimedia information by considering not only the document's text and images data but also the layout structure which determines how a given text block is related to a particular image or video. For example, a Web news page about "Health Care Reform in America" is composed by text describing some event (e.g., Final Senate vote for the reform plans, Obama signs the reform agreement), images (e.g., images about various government involvements over decades) and videos (e.g. Obama's speech video about the decisions) containing additional information regarding the real extent of the event or providing evidence corroborating the text part.

Current state-of-the-art information fusion approaches can be divided into two groups: formal "top-down" methods from the generic knowledge fusion community and quantitative "bottom-up" techniques from the applied Semantic

Web community (Appriou et al., 2001; Gregoire, 2006). Both approaches have their limitations. It will be beneficial to combine both types of approaches so that the fusion decision can be made depending on the type of problem and the amount of domain information it possesses. Saggion et al. (2004) described a multimedia extraction approach to create composite index from multiple and multi-lingual sources. Magalhaes et al. (2008) described a semantic similarity metric based on key word vectors for multi-media fusion. Iria and Magalhaes (2009) exploited information across different parts of a multimedia document to improve document classification. It is important to go beyond key words and attempt representing the documents by the semantic facts identified by IE.

One possible solution is to exploit the linkage information. Specifically, coreference resolution methods should be applied to four types of cross-media data: (1) between the captions of images and context texts; (2) detecting HTML cross-media associations and quantifying the level of image and text block correlation (3) between the texts embedded in images and context texts; (4) between the transcribed texts from the speech in video clips (via automatic speech recognition) and context texts. We can apply a similarity graph to incorporate virtual linkages. For example, when we see images of two web documents containing the same object, we can raise our confidence that such documents are semantically correlated even if the two web documents are from different sources.

4.2 Uncertainty Reduction

When we combine information from images and their associated texts (e.g. meta-data, captions, surrounding text, transcription), one of the challenges lies in the uncertainty of text representation. Therefore it is important to study both how to learn good models from different sources with different kinds of associated uncertainty, and how to make use of these, along with their level of uncertainty in supporting coherent decisions, taking into account characteristics of the data as well as of its source.

The descriptions are usually generated by humans and thus are prone to error or subjectivity. The images, especially the web images, are typically labeled by different users in different languages and cultural backgrounds. It is unrealistic to expect descriptions to be consistent. In speech conversations, many facts are often embedded in questions such as "*It's OK to put Democratic career politicians at the Pentagon and the Justice Department if they're Democrats but not if they're Republicans, is that right?*" This challenge can be generally addressed by strengthening semantic attribute classification methods for Modality, Polarity and Genericity. And if the data sources are comparable, a more direct method of committee-based voting can also be exploited.

However, the fusion process may itself cause data uncertainties. We can follow the co-training framework as described in section 3.1 to reduce uncertainty in fusion. To handle the missing labels, a promising approach is to use graph-based label propagation (Deshpande et al., 2009), which can capture complex uncertainties and correlations in the data in a uniform manner. It's also worth importing the multi-dimensional uncertainty analysis framework described in data mining community (Aggarwal, 2010). The multi-dimensional uncertainty analysis method exactly suits the multi-media fusion needs: it allows us to combine first-order logic with probabilities, modeling inferential uncertainty about multiple aspects - both the context of facts and intended meanings.

4.3 Joint Modeling

IE is generally applied on top of machine generated transcription and automatic structuring that suffer from errors compared to the true content of relations and events. In the context of information fusion we can divide the problem of adaptation into two types: (1) radical adaptation such as from newswire to biomedical articles; (2) modest adaptation such as from newswire to wikipedia or automatic speech recognition (ASR) output. (1) requires a great deal of new development such as ontology definition and data annotation; while (2) can be partially addressed during the information fusion process.

For example, while dealing with speech input, IE systems need to be robust to the noise introduced by earlier speech processing tasks such as ASR, sentence segmentation, salience detection and and speaker identification. Some earlier work (Makhoul et al., 2005; Favre et al., 2008)

showed that using an IE system trained from newswire, the performance degrades notably when the system is tested on automatic speech recognition output. But no general solutions have been proposed to address the genre-specific challenges for speech data.

More specifically, pronoun resolution is one of the major challenges (Jing et al., 2007). For example, in wikipedia a lot of pronouns may refer to the entry entity; while in speech conversation we will need to resolve first and second person pronouns based on automatic speaker role identification; and improve cross-sentence third pronoun resolution by exploiting gender and animacy knowledge discovery methods.

The processing methods of text and other media are typically organized as a pipeline architecture of processing stages (e.g. from pattern recognition, to information fusion, and to summarization). Each of these stages has been studied separately and quite intensively over the past decade. It's critical to move away from approaches that make chains of independent local decisions, and instead toward methods that make multiple decisions jointly using global information. Joint inference techniques (Roth and Yih, 2004; Ji et al., 2005; McCallum, 2006) can transform the integration of multi-media into a benefit by reducing the errors in individual stages. In doing so, we can take advantage (among other properties) of the coherence of a discourse: that a correct analysis of a text discourse reveals a large number of connections from the image information in its context, and so (in general) a more tightly connected analysis is more likely to be correct. For example, prior work has demonstrated the benefit of jointly modeling name tagging and n-best hypotheses, ASR lattices or word confusion networks (Hakkani-Tür et al., 2006).

5 Conclusion

In the current information explosion era, IE technology is facing new challenges of dealing with heterogeneous data sources from different documents, languages and media which may contain a multiplicity of aspects on particular entities, relations and events. This new phenomena requires IE to perform both traditional lower level processing as well as information fusion of factual data based on implicit inferences. This paper investigated the issues of information fusion on a massive scale and the challenges have not been discussed in previous work. We specified the requirements and possible solutions for various dimensions to perform information fusion. We also overviewed some recent work to demonstrate how these goals can be achieved.

The field of information fusion is relatively new; and the nature of different data sources provides new ideas and challenges which are not present in other research. While much research has been performed in the area of data fusion, the context of automatic extraction provides a different perspective in which the fusion is performed in the context of a lot of uncertainty and noise. This new task will provide connections between NLP and other areas such as data mining and knowledge discovery. The progress on this task would save, anybody concerned with staying informed, an enormous amount of time. These are certainly ambitious goals and require long-term development of fusion and adaptation methods. But we hope that this outline of the research challenges will bring us closer to the goal.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. 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. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Charu Aggarwal. 2010. On Multi-dimensional Sharpening of Uncertain Data. *SIAM: SIAM Conference on Data Mining (SDM10).*

A. Appriou-, A. Ayoun, et al. 2001. Fusion: General concepts and characteristics. *International Journal of Intelligent Systems* 16(10).

Javier Artiles, Julio Gonzalo and Satoshi Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. *Proc. Semeval-2007.*

Amit Bagga and Breck Baldwin. 1999. Cross-document Event Coreference: Annotations, Experiments, and Observations. *Proc. ACL1999 Workshop on Coreference and Its Applications*.

K. Balog, L. Azzopardi, M. de Rijke. 2008. Personal Name Resolution of Web People Search. *Proc. WWW2008 Workshop: NLP Challenges in the Information Explosion Era (NLPIX 2008).*

Michele Banko, Michael J Cafarella, Stephen Soderland and Oren Etzioni. 2007. Open Information Extraction from the Web. *Proc. IJCAI 2007*.

Alex Baron and Marjorie Freedman. 2008. Who is Who and What is What: Experiments in Cross-Document Co-Reference. *Proc. EMNLP 2008*.

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. *Proc. ACL-HLT 2008*.

Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-training. *Proc. of the Workshop on Computational Learning Theory.* Morgan Kaufmann Publishers.

T. Brants and R. Stolle. 2002. Finding Similar Documents in Document Collections. *Proc. LREC Workshop on Using Semantics for Information Retrieval and Filtering.*

Yunbo Cao, Hang Li and Li Lian. 2003. Uncertainty Reduction in Collaborative Bootstrapping: Measure and Algorithm. *Proc. ACL 2003*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised Learning of Narrative Schemas and their Participants. *Proc. ACL 09.*

Zheng Chen and Heng Ji. 2009. Can One Language Bootstrap the Other: A Case Study on Event Extraction. *Proc. HLT-NAACL Workshop on Semi-supervised Learning for Natural Language Processing*. Boulder, Co.

Zheng Chen, Heng Ji and Robert Harallick. 2009. A Pairwise Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution. *Proc. RANLP 2009 workshop on Events in Emerging Text Types.*

Amol Deshpande, Lise Getoor and Prithviraj Sen. 2009. Graphical Models for Uncertain Data. *Managing and Mining Uncertain Data (Edited by Charu Aggarwal).* Springer.

Doug Downey, Oren Etzioni, and Stephen Soderland. 2005. A Probabilistic Model of Redundancy in Information Extraction. *Proc. IJCAI 2005*.

Vladimir Eidelman. 2008. Inferring Activity Time in News through Event Modeling. *Proc. ACL-HLT 2008*.

Gunes Erkan and Dragomir R. Radev. 2004. LexPageRank: Prestige in multi-document text summarization. *Proc. EMNLP 2004.*

Oren Etzioni, Kobi Reiter, Stephen Soderland and Marcus Sammer. 2007. Lexical Translation with Application to Image Search on the Web. *Proc. Machine Translation Summit XI*.

Benoit Favre, Ralph Grishman, Dustin Hillard, Heng Ji, Dilek Hakkani-Tur and Mari Ostendorf. 2008. Punctuating Speech for Information Extraction. *Proc. ICASSP 2008*.

Elena Filatova and Eduard Hovy. 2001. Assigning Time-Stamps to Event-Clauses. *Proc. ACL 2001 Workshop on Temporal and Spatial Information Processing.*

Jenny Rose Finkel, Trond Grenager and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proc. ACL 2005*.

E. Gregoire. 2006. An unbiased approach to iterated fusion by weakening. Information Fusion. 7(1).

Prashant Gupta and Heng Ji. 2009. Predicting Unknown Time Arguments based on Cross-event propagation. *Proc. ACL-IJCNLP 2009.*

Dilek Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, Gokhan Tur. 2006. Beyond ASR 1-Best: Using Word Confusion Networks in Spoken Language Understanding. *Journal of Computer Speech and Language*, Vol. 20, No. 4, pp. 495-514.

Laura Hasler and Constantin Orasan. 2009. Do coreferential arguments make event mentions coreferential? Proc. the *7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*.

Jose Iria and Joao Magalhaes. 2009. Exploiting Cross-Media Correlations in the Categorization of Multimedia Web Documents. *Proc. CIAM 2009*.

H. V. Jagadish, Jason Madar, and Raymond Ng. 1999. Semantic compression and pattern extraction with fascicles. *VLDB*, pages 186–197.

Heng Ji, David Westbrook and Ralph Grishman. 2005. Using Semantic Relations to Refine Coreference Decisions. *Proc. HLT/EMNLP 05.*

Heng Ji and Ralph Grishman. 2008. Refining Event Extraction Through Cross-document Inference. *Proc. ACL 2008.*

Heng Ji. 2009. Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks. *Proc. ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora (BUCC 2009): from parallel to non-parallel corpora.*

Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens and Hermann Ney. 2009a. Name Transla-

tion for Distillation. Book chapter for *Global Automatic Language Exploitation*.

Heng Ji, Ralph Grishman, Zheng Chen and Prashant Gupta. 2009b. Cross-document Event Extraction, Ranking and Tracking. *Proc. RANLP 2009.*

Hongyan Jing, Nanda Kambhatla and Salim Roukos. 2007. Extracting Social Networks and Biographical Facts From Conversational Speech Transcripts. *Proc. ACL 2007*.

A. Klementiev and D. Roth. 2006. Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. *Proc. HLT-NAACL 2006*.

Joao Magalhaes, Fabio Ciravegna and Stefan Ruger. 2008. Exploring Multimedia in a Keyword Space. *Proc. ACM Multimedia*.

Inderjeet Mani, Barry Schiffman and Jianping Zhang. 2003. Inferring Temporal Ordering of Events in News. *Proc. HLT-NAACL 2003.*

John Makhoul, Alex Baron, Ivan Bulyko, Long Nguyen, Lance Ramshaw, David Stallard, Richard Schwartz and Bing Xiang. 2005. The Effects of Speech Recognition and Punctuation on Information Extraction Performance. *Proc. Interspeech*.

Gideon Mann. 2007. Multi-document Relationship Fusion via Constraints on Probabilistic Databases. *Proc. HLT/NAACL 2007*.

Andrew McCallum. 2006. Information Extraction, Data Mining and Joint Inference. *Proc. SIGKDD*.

Paul McNamee and Hoa Dang. 2009. Overview of the TAC 2009 Knowledge Base Population Track. *Proc. TAC 2009 Workshop*.

Gilad Mishne and Maarten de Rijke. 2006. Capturing Global Mood Levels using Blog Posts. *Proc. AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.*

Vincent Ng. 2009. Graph-Cut-Based Anaphoricity Determination for Coreference Resolution. *Proc. HLT-NAACL 2009.*

Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. *Proc. WWW.*

Siddharth Patwardhan and Ellen Riloff. 2009. A Unified Model of Phrasal and Sentential Evidence for Information Extraction. 2009. *Proc. EMNLP.*

Matt Richardson and Pedro Domingos. 2006. Markov Logic Networks. *Machine Learning*. 62:107-136.

Ellen Riloff, Charles Schafer, and David Yarowsky. 2002. Inducing Information Extraction Systems for New Languages via Cross-Language Projection. *Proc. COLING 2002*.

Alan Ritter; Stephen Soderland; Doug Downey; Oren Etzioni. 2009. It's a Contradiction – no, it's not: A Case Study using Functional Relations. *Proc. EMNLP 2009.*

Dan Roth and Wen-tau Yih. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. *Proc. CONLL2004*.

Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O., and Wilks, Y. 2004. Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project. *Data Knowlege Engineering*, 48, 2, pp. 247-264.

Roser Saurı and Marc Verhagen and James Pustejovsky. 2006. Annotating and Recognizing Event Modality in Text. *Proc. FLAIRS 2006*.

Len Seligman, Peter Mork, Alon Halevy, Ken Smith, Michael J. Carey, Kuang Chen, Chris Wolf, Jayant Madhavan and Akshay Kannan. 2010. OpenII: An Open Source Information Integration Toolkit. *Proc. the 2010 international conference on Management of data.*

Charles Sutton and Andrew McCallum. 2004. Collective Segmentation and Labeling of Distant Entities in Information Extraction. *Proc. ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields.*

Roman Yangarber. 2006. Verification of Facts across Document Boundaries. *Proc. International Workshop on Intelligent Information Access*.

Alexander Yates and Oren Etzioni. 2009. Unsupervised Methods for Determining Object and Relation Synonyms on the Web. *Journal of Artificial Intelligence*. Res. (*JAIR*) 34: 255-296.

Xiaoxin Yin, Jiawei Han and Philip S. Yu. 2008. Truth Discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowledge and Data Eng.*, 20:796-808.