RC23128 (W0402-147) February 27, 2004
Computer Science

IBM Research Report

Multilingual Annotation of a Shallow Ontology of Entities, Events, and Relations

Judith G. Hochberg, Nanda Kambhatla, Salim Roukos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Multilingual Annotation of a Shallow Ontology of Entities, Events, and Relations

Judith G. Hochberg, Nanda Kambhatla, Salim Roukos
IBM T.J. Watson Research Center
1101 Kitchawan Road Route 134
Yorktown, NY 10598
{nanda,roukos}@us.ibm.com

Abstract

This paper describes an effort to annotate entities, events, and relations in multilingual news stories. We iteratively refined a flat and shallow ontology to achieve fast and consistent annotations that capture more of the essence of news stories than previous work. Initial results obtained by training automatic extractors with the annotated data are promising. We present these results and discuss the challenges we faced.

1 Introduction

In this paper, we describe the creation and annotation of a shallow ontology of entities, events, and relations in multilingual news stories. For us, ***entities*** are objects or abstractions, ***events*** are happenings, and ***relations*** capture explicitly stated relationships among entities and events. Our goals are: creating automatic extractors of entities, events, and relations that capture more of the semantics in news stories than previous work, extracting these from large text collections, and populating a knowledge base with the extracted information. These knowledge bases can be used by diverse applications such as biography extraction, summarization of events, question answering (e.g. to answer queries such as "Which African countries did George W. Bush visit in 2002?"), etc. The output of the annotation is used to train and test automatic extractors.

Several previous efforts at building resources to support automatic extraction, including the MUC evaluations (Chinchor 1998), the CONLL evaluations (CONLL 2003), etc., have focused mostly on named entities of only a few semantic types. Though the MUC evaluations evaluated event and relation extraction, the design of the types (slots or roles) to be extracted was intentionally domain specific. So far, the CONLL evaluations have explored noun phrase chunking, clause detection and named entity extraction. In contrast, the CONLL 2004 evaluation evaluates the extraction of semantic roles in English of some target verbs in propositions.

The Automatic Content Extraction (ACE) evaluations evaluate the extraction of entities, events, and relations in multilingual news stories. The ontology and corpora developed by the Linguistic Data Consortium to support the ACE evaluations (Strassel *et. al*. 2003) are domain independent and capture more of the "who did what to whom" semantics in news stories than previous work.

> Two ***explosions*** were heard near Japan's Defense Ministry late Tuesday and police said ***they*** might have been caused by radicals opposed to the dispatch of Japanese troops to Iraq.
>
> Police said they found two steel pipes that appeared to have been used to launch projectiles from the grounds of a temple near the ministry.
>
> A Defense Ministry official said he was unaware of any injuries or damage at the ministry, but that security officials were still making checks. The ***explosions*** occurred at about 11 p.m.

Figure 1. A fragment of a news story annotated with mentions of some entities and events. All underlined words or phrases are mentions. All mentions in ***bold*** refer to the explosions in Japan (an EVENT-VIOLENCE). Two of the relations in the fragment are timeOf (explosions, 11p.m.) and locatedNear (explosions, Defence Ministry).

Our goal is to significantly improve upon the prior work in the depth of semantics of news stories that our ontology covers. At the same time, we strive for fast and consistent annotation of data using the ontology. Though some of our entity and relation categories were inspired by ACE and other evaluations, we now tag a substantially larger set of entities, events, and relations than LDC, while achieving a fast annotation with inter-annotator agreements comparable to LDC.

The paper is organized as follows. Section 2 describes the ontology of entities, events, and relations that we developed. Section 3 describes the annotation process and initial results with automatic extractors we built using the annotated data. In section 4, we discuss some of our design choices and present our conclusions.

2 Developing the ontology

Our goal is to develop an ontology of entities, events, and relations that can capture the essense of news stories. Our interest is not in capturing *all* the semantic nuances in stories, which might lead to a large and unwieldy set of categories and hence, poor inter-annotator agreement. At the same time, we strive to avoid overly broad categories.

In text, ***mentions*** are the words and phrases that are references to entities and events of interest, and *coreference* is the correspondence between different mentions referring to the same entity or event. Figure 1 shows a news snippet annotated with mentions of some entities, events, and relations among them.

2.1 Entities

Entities are objects or abstractions. We designed 36 entity categories that we iteratively refined to ensure dense and consistent annotations. The final set of entity categories (see Table 1) reflect entities commonly discussed in news stories. Most of the categories were originally developed as potential categories of answers for a question answering system (Ittycheriah *et. al.* 2001).

Table 1: Entity Categories

Category	Examples
People and their properties	
PERSON (singular)	*John Smith, lawyer*
PEOPLE	*Korean, the Petersons*
AGE	*50 years old*
DISEASE	*Hodgkin's disease*
OCCUPATION	*lawyer, president*
SALUTATION	*Ms., Rev.*
Organizations and their properties	
ORGANIZATION	*M.I.T., O'Reilly*
COMPANYROLE	*brokerage house*
Locations	
AREA	*Third World, North America*
ATTRACTION	*Disney World, Monterey Bay*
COUNTRY	*Spain, U.S*
FACILITY	*Lincoln Tunnel, Frick Museum*
GEOLOGICALOBJ	*Mediterranean, Gobi Desert*
LOCATION	*New York, Brooklyn, 5th Avenue*
Dates and Times	
DATE	*"November 2, 2001", 60's*
DATEREF	*Last week, 2 years ago*
TIME	*6:00, 6 pm*
TIMEREF	*last night, late in the evening*
DURATION	*3 hours,* within *several years*
Numbers	
CARDINAL	*3, three, several, hundreds*
ORDINAL	*First, second, third*
PERCENT	*50 percent*
MONEY	*one million dollars*
MEASURE	*4 miles, 4 grams, 4 degrees*
Man-made	
LAW	*Treaty of Guadalupe Hidalgo*
PRODUCT	*Windows, TOEFL*
TITLE_WORK	*Moby Dick, The Lion King*
VEHICLE	*Ford, Corolla, Cessna*
WEAPON	*gun, bomb*
Nature	
ANIMAL	*lion, Smokey the Bear*
FOOD	*banana cream pie*
ORGAN(body part)	*liver, hand, head*
PLANT	*oak*
SUBSTANCE	*boron, ricin*
WEATHER (rare)	*El Niño*

2.2 Events

An event is something that happens at a specific time and place; its scale can be small (a fistfight) or large (a war). We developed a set of 12 event categories. While these categories are not broad enough to cover all events in the world (or even in newspaper stories!), they provide good coverage of key events that are described in news stories. In fact, most stories describe events and entities that played a role in the events, as shown in Fig. 2. As with entity categories, we iteratively refined the event categories. Table 2 shows the final list of event categories.

Albanian *wrecks* two Italian police cars in *chase*
Clinton *urges* companies to *hire* people off welfare
U.S. plane again *attacks* Iraqi mobile missile site
Arrow to *acquire* Premier Farnell business

Figure 2. Sample headlines showing main story events

Table 2: Event Categories	
Category	**Examples**
BUSINESS	loans, acquisitions, bankruptcy
COMMUNICATION	report
CUSTODY	arrest
DEMONSTRATION	rally, parade, strike
DISASTER	train wreck, earthquake
LEGAL	lawsuit, hearing, execution
MEETING	meeting, conference
PERFORMANCE	show, graduation, dedication
PERSONNEL	resignation, hiring
SPORTS	game, tournament
VIOLENCE	battle, war, murder
GENERIC	economic crisis, death

2.3 Relations

Relation categories capture relationships among entities and events. We have defined 32 relation categories. Several of these were inspired by the ACE relation categories (Strassel *et. al*. 2003). However, we extended and modified the ACE categories to cover events and entities not marked by ACE. Table 3 lists and exemplifies the relation categories, broken down into some general types according to their argument restrictions. (The possible arguments of a relation category are part of its definition; for example, a person can be a citizenOf a COUNTRY only, and not of an AREA, LOCATION, or other locational entity.)

Table 3: Relation Categories	
Principal argument types	**Relations/Examples**
events	affectedBy (town, plague) agentOf (killer, murder) participantIn (player, game) instrumentOf(gun, shooting) topicOf (meeting, war) timeOf (meeting, Monday) locatedAt (meeting, office)
person, person	relative (her, mother) colleague (John, teammate) playsRoleOf (actor, Nixon)
person, org	managerOf (Bush, USA) memberOf (teller, bank) founderOf (Edison, GE)
person/org, location	locatedAt (Bangalore, India) near (Pakistan, India) residesIn (Carter, Georgia) citizenOf (Kennedy, USA) basedIn (Dell, Texas)
person/org, various	hasProperty (Paul, Pope) hasProperty (Chase, bank) hasProperty (him, AIDS)
	diedOf (she, cancer)
person, date	bornOn (Washington, 1732) diedOn (Washington, 1799)
org, org	partner (bank, Bank One) competitor (they, company) partOf (half.com, ebay)
person, person/org	clientOf (person, lawyer) spokespersonFor (Smith, GE)
title_work, person/org	awardedBy (Oscar, Academy) awardedTo (Oscar, Jackson) authorOf (Melville, book)
singular, plural	partOfMany (dog, dogs)
cardinal, location	populationOf (300, town)
person/org, various	ownerOf (my, car) ownerOf (France, missile)

Of the categories listed in Table 3, those whose names lack a preposition are symmetric; e.g., since France is *near* Spain, Spain is *near* France. Three categories (*locatedAt*, *partOf*, and *partOfMany*) are transitive; e.g., since Miyun is in Beijing and Beijing is in China, then Miyun is in China.

3 The KDD corpus: annotation of entities, events, and relations

We have created a corpus (called the *KDD corpus*) of documents annotated with the entities, events, and relation categories described above, by sevaral native speakers of English, Chinese and Arabic. Our annotators did not have any formal linguistics background. New annotators went through a training period, as they absorbed the definitions of our categories. Annotators were encouraged to actively seek clarifications when in doubt using instant messaging and other means of communication.

For entities, we annotate the head of named (e.g. “George Bush”), nominal (e.g. “president”) and pronominal (e.g. “he”) mentions of entities. We mark only mentions referring to specific entities, avoiding mentions of generic entities like "*Man” in “Man* is a foolish *animal*". We annotate long noun phrases as sequences of mentions like “*[high school][football team]* “and “*[US][vice president][Dick Cheney]”*.

Metonomy occurs when a mention of an entity is used to refer to another entity that can be of a different semantic type, e.g. in “Washington announced that …”, “Washington” is being used to refer to the U.S. government. In such cases, we mark the intended semantic type (ORGANIZATION for “Washington” in the previous example) of the mention.

For events, we tag verbs or nominalizations that serve as anchors of the main events in a story. We ignore minor sub-events (e.g. we tag a fight but not individual punches). The goal is to capture the essence

rather than to tag every verb. As with entities, we also annotate the coreference of multiple mentions of the same event within a story.

We chose to annotate only those relations that were supported by explicit textual evidence. Thus we would tag Queen Elizabeth as a *relative* of Prince Charles (and vice versa) only if a document explicitly stated the fact. For each relation, annotators mark the time when the relation is true: e.g. present, past, future, hypothetical, etc. with respect to the time of writing of the document. Currently, we only tag relations within a sentence.

We used the Alembic toolkit developed by Mitre (Day et. al. 1997) for annotating mentions of entities and events, and the coreference of mentions. For annotating relations, we used an in house tool that enforced the relation argument restrictions discussed in section 2.3.

For each document, a single annotator marked all mentions of entities, all mentions of events, coreference, and relations in that order in four separate passes. We achieved an annotation rate of approximately 17 words per minute (wpm) for annotation of entities, events, and relations.

3.1 The density of annotation

	Entropy (KDD)	***Entropy (ACE)***
Mentions	**3.02**	2.50
Relations	**2.35**	1.98

Figure 3. The entropy of the distribution of mentions and relations in KDD and ACE corpora. The higher entropy for annotations in the KDD corpus indicates a more uniform spread of mentions and relations across entity, event, and relation types.

So far, the KDD corpus comprises over 400K words of English, over 400K characters of Chinese, and over 150K words of Arabic data. For the English documents, on average, there is a mention of an entity or event every 4 words and an instance of a relation every 10 words (equivalent to roughly two relations per sentence). We are interested in the ***density*** of annotation, since it can be a crude measure of the depth of semantics captured by the annotation. The density we achieved compares favorably to LDC's annotation for the ACE evaluations. For the ACE 2002 and 2003 training data (around 330K words), there is a mention of an entity every 6 words and an instance of a relation every 27 words (equivalent to roughly one relation per sentence) on average.

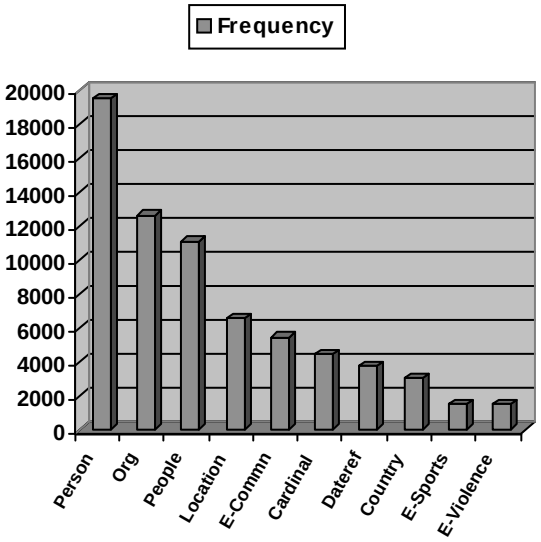

Figure 4. The number of mentions annotated for the ten most frequently annotated entity and event categories.

We computed the ***entropy*** of the distribution of different types of mentions and relations for the ACE corpus (LDC's annotated corpus comprising training data for the 2002 and 2003 evaluations) and the KDD corpus. A higher entropy for mentions would suggest a more uniform distribution of mentions of different entity and event types within a document. Conversely, a lower entropy would indicate a dominance of mentions of a few entity or event types. Figure 3 shows the entropy for mentions and relations annotation for the ACE and KDD corpora. Note that the entropy of the distribution of entity and event mentions is significantly higher for the KDD corpus than for the ACE corpus. Similarly, the entropy of the distribution of relations is significantly higher for the KDD corpus. These results suggest

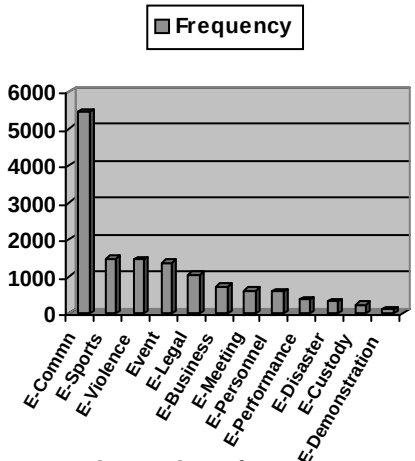

Figure 5. The number of mentions annotated for event categories.

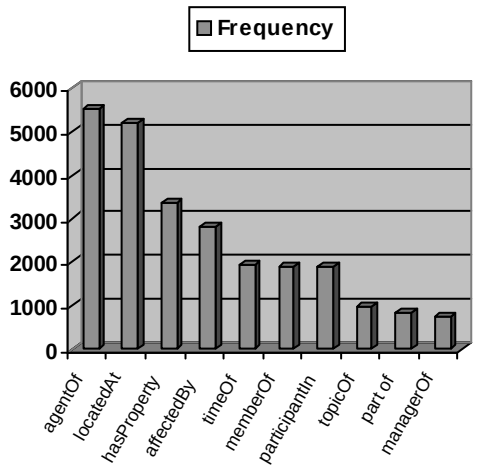

Figure 6. The number of relations annotated for the ten most frequently annotated relation categories.

that our larger set (than ACE) of entity, event, and relation categories are successfully capturing more of the semantics in documents, since on average, the distribution of mention and relation types is more spread out among a larger, broader set of categories.

On average, an entity has 1.7 mentions in an English document, with the number varying greatly based on the category (e.g. on average, a PERSON entity has 3.6 mentions, while a FACILITY entity has 1.4 mentions in a document). Figure 4 shows the number of mentions (in the whole corpus) of the ten most frequently mentioned entity and event categories. Not unexpectedly, our documents (which are news stories) mention persons, organizations, locations, communications, and events of violence much more often than other categories. Figure 5 shows the number of mentions (in the whole corpus) for the event categories. Note that we have five times more mentions of communications events (press conferences, announcements, reports, etc.) than any other event category.

Figure 6 shows the number of instances of annotated relations (in our corpus) of the ten most frequently annotated relation categories. Note that we have a lot more instances of relations between events and entities than between entities.

3.2 Annotation challenges

This annotation process is a difficult one that involves multiple steps, each with its own categories and rules. This subsection discusses three of the most difficult annotation challenges that have arisen during the project. We have tried to reduce inter-annotater error stemming from these challenges as much as possible by defining, documenting, and drilling specific rules for each case.

The first challenge is identifying the head of a mention. The head can be a single word (in English, usually the last token of a noun phrase) or a tightly bound phrase of strongly associated words. For example, in "injured *passengers*", "southern *Spain*", and "every *person*", the first word is a modifier rather than part of the head. In contrast, *auto dealership*, *prime minister*, *economic adviser*, *football team*, and *Green Bay Packers* are all multi-word heads.

The second challenge is coreferencing documents with multiple related entities. Even in a short news story, it can be surprisingly hard to decide how many distinct entities are referred to; see Figure 7 for a complicated but not unusual example.

MAE HONG SON, Thailand - LRB - AP - RRB - : Khun Sa 's soldiers, scattered in Burma 's jungles since the opium warlord surrendered in 1996, are uniting into a 15,000 - man **army**[1] to resume their war against the government in Rangoon, a Thai security official said Wednesday.

The new **guerrilla force**[2] recently acquired a huge cache of weapons and Burma watchers expect a sharp surge in fighting in the coming months, said the official who spoke on condition of anonymity.

During the past three weeks, three ethnic Shan **rebel forces**[3], **some**[4] formerly led by Khun Sa, have been working together and launching coordinated hit - and - run attacks on government troops throughout central and southern Shan State, the official said.

For more than a decade, Khun Sa .... commanded an **army**[5] **that**[6] ranged between 10,000 - 20,000 men and controlled the lion 's share of the area 's traffic in opium, the raw material for heroin.

As time went by, however, rival opium **armies**[7] sprang up, diminishing his power.

Figure 7. How many opium army entities appear in this 174-word story? We count five: Khun Sa's original army (5-6), Khun Sa's old rival armies (7), the three rebel forces (3), the subset of the rebel forces formerly led by Khun Sa (4), and the reconstituted army (1-2). Therefore mentions 1 and 2 should be coreferenced, and likewise mentions 5 and 6.

The third challenge is picking the correct pair of mentions to annotate a relation, in cases where two or more pairs are available. An example is the fourth paragraph of Figure 7, in which "Khun Sa" was managerOf either "army" (mention 5) or "that" (mention 6). In such cases the annotaters choose the mention pair with the best syntactic evidence for the relation – in this case, mention 5, which is a co-argument with "Khun Sa" of the verb "commanded". Recognizing syntactic evidence often amounts to performing an informal parse of a sentence, which can be quite difficult for annotaters without formal linguistic training.

3.3 The quality of annotation

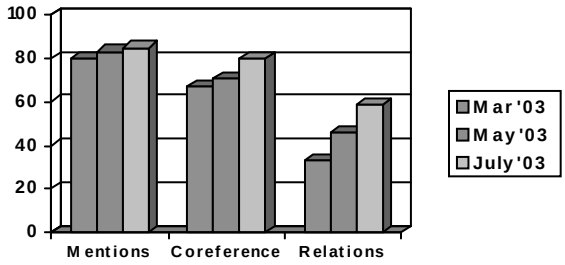

Figure 8. Inter Annotator Agreement (ITA) (measured here as the F-measure) improvement over time.

We had some documents annotated by two different annotators and we compared their agreement with each other. We computed the F-measure (harmonic mean of precision and recall) between the annotations of the two annotators, treating the annotations of one of them as the truth. We separately computed the F-measure for all mentions of entities and events, the coreference between mentions, and relations among mentions.

Figure 7 shows the F-measure obtained for the Inter-Annotator Agreement (ITA) for mentions, coreference and relations for annotation of English documents. We obtained similar scores for annotation of Chinese documents. We havn't yet measured ITA for Arabic documents. Our scores for mentions, coreference and relations are comparable to those obtained by LDC for annotating the ACE corpus[1]. Note that the ITA improved significantly with time as the annotation progressed. This can be attributed to the iterative sharpening of our entity, event, and relation categories to enable consistent annotations and the increasing comfort level of the annotators due to experience.

In figure 7, the ITA is lower for coreference and relations than mentions. The measurement of agreement in coreference annotation reflects the disagreements in mention annotation also. Similarly, the measurement of agreement in relation annotation reflects the disagreements in mention annotation. Thus, if there is only 80% agreement (F-measure) in mention annotation and an 80% agreement in relation types, we would expect an agreement of only about 51% (F-measure) for relations, since agreement for relations implies agreeing on both the mentions which are arguments for the relation and the relation type (0.8 * 0.8 * 0.8 = 0.51).

Relation annotation can be hard, since, intuitively, the decision on the type of relationship (or lack thereof) between two mentions requires more contextual information than the decision on the type of entity or event for a phrase. Overall, our ITA of around 85% for mentions, around 80% for coreference (while inheriting disagreements for mentions), and around 58% (while inheriting disagreements for mentions) are good and comparable to those obtained by similar annotation efforts like ACE (Strassel *et. al.* 2003).

3.4 Initial results with automatic extractors

Subtask	***F-measure (English)***	***F-measure (Chinese)***
Mention Detection	75	82
Mention Coreference	69	70
Relation Extraction	37	34

Figure 9. F-measure obtained by automatic extractors for detecting mentions of entities and events, coreference between the mentions, and relations among them for English and Chinese documents.

We trained separate statistical models for detecting mentions of entities and events, coreference among mentions and relations among them. We separated the annotated data (described in the previous section) into training and test sets and used the training sets to train the statistical models. We used linear and log-linear models as used by (Florian *et. al.* 2004) to train our models.

Figure 9 summarizes the results we obtained for English and Chinese models. As with ITA, note that the inheritance of mention errors for coreference resolution and relation extraction contributes to lowering those scores.

4 Discussion

In developing our ontology, we deliberately tried to keep the categories as simple and intuitive as possible, to enable fast and consistent annotation by people with good reading comprehension skills. At the same time, we wanted to have enough complexity to capture the essence of news stories. One specific goal was to capture more of the semantics of news stories than previous work (e.g. MUC, CONLL, ACE evaluations).

[1] Note that LDC measures inter annotator agreement using the ACE value metric. The reader is referred to the ACE website for more details.

To satisfy these potentially conflicting objectives, we went through several iterations of defining (or redefining) the ontology and annotating some documents with it to measure the density and quality of annotations. To reduce complexity, we decided to eschew hierarchical categories in favor of a flat list. We treated several properties of entities as entities (e.g. age, occupation). Enforcing argument restrictions for each relation also reduced complexity by limiting the number of potential relations for each mention. To promote domain independence, we did not include very specific event templates as in MUC (Chinchor 1998). However, as Figure 1 showed, it is possible to reconstruct event templates from the annotated events, coreference, and relations among them.

We have created a corpus (the KDD corpus) of documents annotated with our ontology of entities, events, and relations. Our results comparing the KDD corpus and the ACE corpus suggest that our ontology captures more of the semantics of documents than the ACE annotations. Our annotation is denser than ACE in both mentions of entities and events and instances of relations. Moreover, the entropy of the distribution of entity and event types and relation types is significantly higher for the KDD corpus.

In summary, we have presented a shallow ontology of entities, events, and relations. We have described the annotation of the KDD corpus with the ontology. The annotation was both fast (17 wpm), consistent (ITA around 80%) and captured more semantic types on average than the ACE corpus. Preliminary results show that it enables the creation of accurate automatic extraction modules. We argue that creating a simple, flat list of semantic categories enabled us to achieve a fast, consistent annotation of the essence of news stories.

References

ACE: The NIST ACE evaluation website : http://www.nist.gov/speech/tests/ace/

Nancy Chinchor, 1998. "Overview of MUC-7." *In Seventh Message Understanding Conference (MUC-7):* Proceedings of a Conference Held in Fairfax, April. Morgan Kaufmann.

CONLL 2003: The Conference on Computational Natural Language Learning 2003 : http://cnts.uia.ac.be/conll2003/ner/

David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson and Marc Vilain. Mixed-Initiative Development of Language Processing Systems. 1997. *ANLP-5 Proceedings*.

Abraham Ittycheriah, Martin Franz and Salim Roukos. 2001. IBM's Statistical Question Answering System. *TREC-10 Proceedings*: 258-264.

Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov and Salim Roukos. 2004. A Statistical Model for Multilingual Entity Detection and Tracking. *HLT-NAACL'2004 Proceedings*.

Stephanie Strassel, Alexis Mitchell and Shudong Huang. Multilingual Resources for Entity Detection. *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, 2003.