Joint Inference for Information Extraction and Translation

Heng Ji
(hengji@cs.nyu.edu)

Advisor: Prof. Ralph Grishman
Computer Science Department
New York University

February, 2008
Outline

- Problem: Performance Limitation of Traditional IE frameworks
- General Solution: A New IE Framework based on Joint Inference
- Concrete Solution
  - Linguistic Evidence of Component Interactions
  - Re-ranking Algorithms to Incorporate Component Interactions
- Case Studies
  - Coreference/Relation/Event as Feedback to Improve Name Tagging
  - Relation as Feedback to Improve Coreference Resolution
  - Entity Translation as Feedback to Improve Source Language Entity Extraction
- Related Work
- Conclusion and Future Work
- My Research Overview
What is Information Extraction (IE)?
IE Example

- I'd like to fly on Northeast Airlines, and uh fi-fly from New York to Pittsburgh, the exciting city located in Pennsylvania.
- Um uh does Northeast have a flight between five P M and six P M.
- W- wha- what type of questions would I get during my interview talk by the way.

<table>
<thead>
<tr>
<th>Origin</th>
<th>New York</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination</td>
<td>Pittsburgh</td>
</tr>
<tr>
<td>Agent</td>
<td>Northeast Airlines</td>
</tr>
<tr>
<td>Time-Within</td>
<td>five P M</td>
</tr>
</tbody>
</table>
IE Components

- Name/Nominal Mention Tagger
  - "New York"
  - "Pittsburgh"
  - "Pennsylvania"
  - "Northeast"
  - "city"

- Coreference Resolver (Entity Clustering)
  - "Northeast Airlines" = "Northeast"
  - "Pittsburgh" = "city"

- Relation Tagger
  - "city" is located in "Pennsylvania"

- Event Tagger
  - "New York" is the origin of the movement event triggered by "fly"

Automatic Content Extraction (ACE) IE
- Entity types: 7 types (person, geo-political, organization, location, facility, weapon, vehicle)
- Relation types: 6 types, 18 subtypes (organization-affiliation, personal-social, located, …)
- Event types: 8 types, 33 subtypes (movement, business, contact, …)
Problem

Performance Limitation of Traditional IE Frameworks
Sequential IE Framework

Errors are compounded from stage to stage

Precision: 100% 90% 80% 70% 60%
Monolithic IE Framework

More Independent Design?

Monolithic IE Model by a wide feature space, no clear interactions between components

Precision

- names: 65%
- coreference: 65%
- relation: 65%
- event: 65%

Still not satisfactory
Goal: Design a New IE Framework

- Question: Can we combine the advantages of two traditional frameworks while overcoming their weaknesses?

- Answer: Retain sequential processing modules but capture interactions between different components
General Solution

Design a New IE Framework based on Joint Inference
A New IE Framework based on Joint Inference

Stage_1 \rightarrow \text{Input} \rightarrow \text{Baseline} \rightarrow \text{Multiple Hypotheses} \rightarrow \text{Feedback Knowledge Encoder} \rightarrow \text{Hypothesis Selector} \rightarrow \text{Best Hypothesis}

Source IE

Stage_{j+1} \rightarrow \text{MT} \rightarrow \text{Stage}_s

Symbiosis
IE Framework based on Joint Inference

1. Keep sequential design

2. Each stage generates multiple hypotheses and propagates them to subsequent stages

3. Use feedback information from subsequent stages to correct, re-score or re-rank these multiple hypotheses

4. The new top hypothesis is the final extraction result
Concrete Questions

1. What kind of useful linguistic interactions can be used?

2. How can we select/organize these types of interactions into inference rules or machine learning models?
Concrete Answer #1

Types of Linguistic Interactions
Sources of Interaction Overview

- A natural discourse is more likely to be cohesive, to have name candidates which are linked by coreference and semantic relations, or involved in certain events.

- Semantic relations and events impose type constraints (or preferences) on their name arguments.

- Coreference is by definition a semantic relation; so feedback from relation detection can provide deeper evidence.

- Language-specific information in the target language from machine translation output can provide useful feedback to source information extraction.
Example: Name Tagging from an IE Interaction View

- \( i^{th} \) Input Sentence
- Baseline Name Tagger
- N-Best Hypotheses
- Relation Tagger
- Event Tagger
- Coreference Resolver
- Feedback Feature Extractor
- Re-Ranker
- Best Hypothesis

Flowchart:

1. Input Sentence
2. Baseline Name Tagger
3. N-Best Hypotheses
4. Relation Tagger
5. Event Tagger
6. Coreference Resolver
7. Feedback
8. Re-Ranker
9. Best Hypothesis
The previous president Milosevic was beaten by the Opposition Committee candidate Kostunica.

Milosevic was born in Serbia’s central industrial city.
Take Wider Context

- **jth Input Sentence**
  - **Baseline Name Tagger**
  - **N-Best Hypotheses**
    - **Feedback Feature Extractor**
      - **Re-Ranker**
        - **Best Hypothesis**

- **Sent 0**: 
  - **Sent 1**: 
  - **Sent i**: 
  - **Cross-Sentence**
  - **Cross-Stage Feedback**

- **Relation Tagger**
  - **Event Tagger**
    - **Coreference Resolver**
Even Wider Context

- \(i^{th}\) Input Sentence
- Baseline Name Tagger
- N-Best Hypotheses
- Feedback Feature Extractor
- Re-Ranker
- Best Hypothesis
- Relation Tagger
- Event Tagger
- Cross-doc Coreference Resolver

Even Wider Context...
Concrete Answer #2

Algorithms for Incorporating Various Interactions
Algorithms Overview

- **Rules (Case Study #3 later)**
  - Convert interactions into post-processing rules
  - Use rules to adjust or correct the sequential baseline outputs

- **Supervised Re-Ranking (Case Study #1 and #2 later)**
  - Baseline generates N-Best hypotheses and initial ranking/confidence
  - Build a second-phase supervised ranking model
  - Encode the baseline output confidence and component interaction knowledge as features
  - Re-predict new rankings for the hypotheses, generate the new top hypothesis as the final output
  - Details next…
Re-Ranking: Hypothesis Representation

- Ranking: to accurately rank a set of objects

- The choice of “objects” depends on the goal
  - Each hypothesis represents an alternative result for each document/sentence/entity/mention, etc.
  - Use “gold standards”: F-measure for name tagging/coreference resolution, ACE metric to measure entity extraction, etc.
  - Example: To get best name tagging results: set each object as (sentence-based) name hypothesis
Name Hypothesis Example

- **Sentence**
  
  Slobodan Milosevic was born in Serbia.

- **Generate N-Best hypotheses**
  
  Hypo0: Slobodan Milosevic was born in <PER>Serbia</PER>.
  Hypo1: <PER>Slobodan</PER> Milosevic was born in <GPE>Serbia</GPE>.
  Hypo2: <PER>Slobodan Milosevic</PER> was born in <GPE>Serbia</GPE>.
  ...

- **Goal: Re-Rank hypotheses and get the new best one**
  
  Hypo0’: <PER>Slobodan Milosevic</PER> was born in <GPE>Serbia</GPE>.
  Hypo1’: <PER>Slobodan</PER> Milosevic was born in <GPE>Serbia</GPE>.
  Hypo2’: Slobodan Milosevic was born in <PER>Serbia</PER>.
  ...

23/56
Re-Ranking: Learning Function $f$

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Score</th>
<th>Score based Direct Re-Ranking</th>
<th>Classification based Direct Re-Ranking</th>
<th>Pair wise-Comparison based Indirect Re-Ranking</th>
<th>Baseline-Comparison based Indirect Re-Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>80%</td>
<td>$f(h_1) = 0.8$</td>
<td>$f(h_1) = -1$</td>
<td>$f(h_1, h_2) = -1$</td>
<td>$f(h_2) = 1$</td>
</tr>
<tr>
<td>$h_2$</td>
<td>90%</td>
<td>$f(h_2) = 0.9$</td>
<td>$f(h_2) = 1$</td>
<td>$f(h_1, h_3) = 1$</td>
<td>$f(h_2) = 1$</td>
</tr>
<tr>
<td>$h_3$</td>
<td>60%</td>
<td>$f(h_3) = 0.6$</td>
<td>$f(h_3) = -1$</td>
<td>$f(h_1, h_3) = 1$</td>
<td>$f(h_3) = -1$</td>
</tr>
</tbody>
</table>
Re-Ranking: Learning Model

- **MaxEnt-Rank**
  - Log-linear classifier, incorporate rich features automatically, compute a reliable ranking probability for each hypothesis
  - This work used OpenNLP (Generative Iterative Scaling for training), and pair-wise comparison based indirect ranking
  - (Charniak and Johnson, 2005) use MaxEnt-Rank for Parsing

- **SVM-Rank**
  - Pair-wise comparison based indirect ranking
  - This work used SVMLight (Joachims, 1998) with linear kernel
  - (Shen and Joshi, 2003) use SVM-Rank for Parsing and MT

- **p-Norm Push Ranking**
  - AdaBoost-style supervised ranking algorithm (Rudin, 2006), generalizes RankBoost (Collins, 2002)
  - This work used score based direct ranking
We prefer A to B because we only care about the ‘top’ result.

“$p$” represents how much to concentrate at the top of the hypothesis list. If “$p$” is large, then we care more about the rank quality on the top of the hypothesis list (a large “push”); and would rather sacrifice the overall ranking quality.

Which ranking result is ‘better’ (for our task)?

- We prefer A to B because we only care about the ‘top’ result.
- “$p$” represents how much to concentrate at the top of the hypothesis list. If “$p$” is large, then we care more about the rank quality on the top of the hypothesis list (a large “push”); and would rather sacrifice the overall ranking quality.
Case Studies

- Baseline cross-lingual IE system
  1. Coreference/Relation/Event as Feedback to Improve Name Tagging
  2. Relation as Feedback to Improve Coreference Resolution
  3. Entity Translation as Feedback to Improve Source Language Entity Extraction
Baseline Cross-lingual IE System

- Name Tagger: HMM model
- Nominal Tagger: statistical noun phrase chunker and table look-up for head identification
- Coreference Resolver: rules + MaxEnt models
- Relation Tagger: k-Nearest Neighbor algorithm
- Event Pattern Acquisition: parsing + semantic role labeling, extract patterns from ACE data and COMLEX
- Entity Translation: RWTH statistical phrase-based MT
Case Study #1

Using Coreference/Relation/Event to Improve Name Tagging

(Ji and Grishman, COLING/ACL 2006)
Name Re-Ranking Feature Overview

- Prefer the hypothesis including names:
  - Matching name structure constraints better
  - Coreferred more frequently
  - Coreferred by stronger links (name-name > apposition > name-nominal)
  - Matching the entity type constraints in relation pattern better
  - Matching the entity type constraints in event pattern better
  - Getting higher voting rate among N hypotheses
  - And many more…
### Name Re-Ranking Feature Example

| Text 1 | The previous president **Milosevic** was beaten by the **Opposition Committee** candidate **Kostunica**. |
|        | <PERSON>: 2 | <ORGANIZATION>: 1 | <PERSON>: 2 |
| S 11   | Kostunica joined the committee in 1992. |

| Text 2 | **Milosevic** was born in Serbia’s central industrial city. |
| S 21   |

- “CorefNum” Feature: the names in the candidate hypothesis are referred to by CorefNum other mentions
- For the current name hypothesis of S11: CorefNum = 2+1+2=5
Further Improvement: Staged Name Re-Ranking

- Handle cross-sentence features
  - Only 38% of coreference links are within the same sentence
  - Need accurate re-ranking for other sentences to use coreference
  - Coreference helps to globally propagate good name decisions (next slide)
Coreference helps to globally propagate good name decisions

The previous president Milosevic was beaten by the Opposition Committee candidate Kostunica.

Milosevic was born in Serbia’s central industrial city.

(Coreference based Re-Ranker: propagate good decisions)

<PERSON>: Confirmed by Event based Re-Ranker
Name Experiments

- Training data for the baseline and re-rankers are from ACE02~05 training corpora
- Test on 20 English texts (743 names) and 100 Chinese texts (2813 names) from ACE04 training corpora
- Scoring metric
  - $S$: name set returned by the name tagger
  - $K$: key name set
  - Recall = $(S \cap K) / K$
  - Precision = $(S \cap K) / S$
  - F-Measure = $2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision})$
## Name Re-Ranking Performance

<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Baseline</td>
<td>88.0</td>
<td>85.2</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>After Re-Ranking</td>
<td>88.4</td>
<td>87.2</td>
<td>87.8</td>
</tr>
<tr>
<td>Chinese</td>
<td>Baseline</td>
<td>87.4</td>
<td>87.6</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>After Re-Ranking</td>
<td>91.8</td>
<td>90.6</td>
<td>91.2</td>
</tr>
</tbody>
</table>
Performance of Different Name Re-Ranking Models

<table>
<thead>
<tr>
<th>Model for Chinese Names</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Baseline</td>
<td>87.4</td>
<td>87.6</td>
<td>87.5</td>
</tr>
<tr>
<td>Oracle (N=30)</td>
<td>(96.2)</td>
<td>(97.4)</td>
<td>(96.8)</td>
</tr>
<tr>
<td>MaxEnt-Rank</td>
<td>91.8</td>
<td>90.6</td>
<td>91.2</td>
</tr>
<tr>
<td>SVMRank</td>
<td>89.5</td>
<td>90.1</td>
<td>89.8</td>
</tr>
<tr>
<td>p-Norm Push Ranking</td>
<td>91.2</td>
<td>90.8</td>
<td>91.0</td>
</tr>
</tbody>
</table>
Case Study #2

Using Relation to Improve Coreference Resolution

(Ji et al., HLT/EMNLP 2005)
But the unknown culprits were not able to view or steal the company's crucial source code for its Windows or Office software, a company spokesman said Friday afternoon.

Speaking earlier to Microsoft programmers and reporters at a seminar in Stockholm, Sweden, Steven Ballmer, the company's chief executive, said, "It is clear that hackers did see some of our source code."
The Texas governor and Republican nominee was described as in good spirits, but disappointed after Friday's Florida Supreme Court order for an immediate recount of so-called undervotes missed in machine tallies.

But Bush officials did little to hide their obvious dismay with Friday's 4-3 state Supreme Court ruling favoring Democrat Al Gore.
Corference Resolution from a Relation View

Pattern (1):
Same_Relation ^ ~CorefA $\Rightarrow$ CorefBLessLikely

Pattern (2):
~Same_Relation ^ CorefA $\Rightarrow$ CorefBLessLikely

Pattern (3):
Same_Relation ^ CorefA $\Rightarrow$ CorefBMoreLikely
Coreference Re-Ranking Features

- Coreference probability from baseline reference resolver
- The type of the applicable relation/coreference pattern
- Relation type
- Relation subtype
- Relation direction
- The reliability of the rule instantiation
- The text of both mentions
Coreference Experiments

- Training data from ACE 02~04 training corpora
- Test on 65 English texts and 100 Chinese texts from ACE04 training corpora
- Scoring metric: MUC F-measure scoring
- Performance

<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Baseline</td>
<td>87.3</td>
<td>77.2</td>
<td>81.9</td>
</tr>
<tr>
<td></td>
<td>After Re-Ranking</td>
<td>87.5</td>
<td>80.3</td>
<td>83.7</td>
</tr>
<tr>
<td>Chinese</td>
<td>Baseline</td>
<td>76.3</td>
<td>75.0</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>After Re-Ranking</td>
<td>76.5</td>
<td>76.1</td>
<td>76.3</td>
</tr>
</tbody>
</table>
Using Entity Translation to Improve Entity Extraction

(Ji and Grishman, RANLP 2007)
Entity Extraction from a Machine Translation View

- Look outside of language: mono-lingual → cross-lingual feedback
- Large aligned bilingual corpora are available; Can IE benefit from it indirectly?
- Use statistical Machine Translation which distills bilingual corpora
- Can we avoid MT noise? → Use confidence estimation
Confidence Estimation by Voting Mentions of an Entity

- Apply within-doc and cross-doc coreference based voting over mentions of an entity

- **Margin Confidence**  = \( \text{Frequency (Best Assignment)} - \text{Frequency (Second Best Assignment)} \)
  
  an ‘assignment’ is an extraction or translation result for each mention

- Propagate the most common extraction and translation with high confidence

- Examples next…
Using Translation to Adjust Source Mention Identification

- Translation is not capitalized?
- Has few mentions coreferred to? (1)
- Confidence is low? (0)

- Fix the mention tag into nominal.
Using Translation to Adjust Source Isolated Mention Boundary

- Translation is not capitalized?
- Boundary is conflicted with names with the best translation?

- Adjust boundary to be consistent with the name with best translation.
Using Translation to Adjust Source Coreference

- Both the best and second best translations have similar high frequency?
- Two mentions overlap in spelling?
- Split the source entity into two entities.
Using Extraction to Adjust Target Entity Translation

- High confidence?
- Translation is different from the best one?

- Replace the translation with the best one.
Entity Extraction and Translation System Architecture

In this work, S= Chinese and T=English
Entity Extraction and Translation Experiments

- Entity Translation is trained from LDC bilingual corpora (600M words) and GigaWord
- Test on 67 texts from ACE07 training corpora, including 2077 names and 1907 entities
- Scoring: name F-measure and entity ACE value
- Performance

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>Baseline</th>
<th>After Joint Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name Tagging F-Measure</td>
<td>87.3</td>
<td></td>
<td>89.2</td>
</tr>
<tr>
<td>Entity Extraction ACE Value</td>
<td>66.6</td>
<td></td>
<td>67.9</td>
</tr>
<tr>
<td>Entity Translation ACE F-Measure</td>
<td>35.1</td>
<td></td>
<td>38.3</td>
</tr>
</tbody>
</table>
Related Work about Component Interaction

- Interaction between Name Tagging and Coreference Resolution
  - Zhou et al., 2005
  - Daume III, 2006
  - Poon and Domingos, 2007
  - Magdy et al., 2007

- Interaction between Name Tagging and Relation Detection
  - Roth and Yi, 2004
    - Linear Programming, Name Classification Only
  - Yangarber and Jokipii, 2005

- Interaction between Coreference Resolution and Semantic Relation
  - This Work
  - Bean and Riloff, 2004
  - Ponzetto and Strube, 2005
  - Ng, 2007
  - Jing et al., 2007

- Interaction between Information Extraction and Machine Translation
  - Huang and Vogel, 2002
  - Require aligned bilingual corpora
  - Florian et al., 2007
  - Expand IE Training data by MT

- Interaction between Name Tagging and Coreference Resolution

- Interaction between Name Tagging and Relation Detection

- Interaction between Coreference Resolution and Semantic Relation

- Interaction between Information Extraction and Machine Translation
Conclusion

- Proved an N-Best strategy based on component interactions is effective for IE
- Multiple stages are benefit rather than burden: stages sharing broader and deeper knowledge
- Built interaction between IE and MT
- Applied various ranking techniques to IE
- Simple, low-cost, language/baseline-independent
Research Plan: IE Adaptation for Biomedical Domain

- Motivation: Apply IE to the domains which have a direct impact on people’s lives

- Testing a state-of-the-art newswire IE system on medical texts
  - Name tagging
    - ~90% for newswire vs. ~10% for medical text;
    - reason: many unknown capitalized words such as “ED”, “EKG”, “ES”
  - Coreference Resolution
    - ~80% recall for newswire vs. ~20% for medical text;
    - reason: the anaphor is too far away from the antecedent. e.g. five sentences between “female” and “she”
  - Event Extraction
    - ~60% for newswire vs. 30% for medical text;
    - reason: different word senses. is “left sternal area” a “movement” event?

- Plan
  - Semi-supervised learning for domain adaptation
  - Incorporate medical domain ontology into IE system
  - Re-optimize machine learning parameters for medical domain
Other Plans for Medical NLP

- Confidence Estimation based Uncertainty and Negation Tagging
  - Machine learning based inference, using global context
  - Use NYU FactBank

- Using Deep Knowledge in Learning Algorithms for Coreference Resolution
  
  One woman marched with the words "shock and awe" daubed across her back in red ink, a reference to the bombing campaign unleashed Friday on Baghdad.

  Liana Owen drove 10 hours from Pennsylvania to attend the rally in Manhattan with her parents. "It's important that people all over the world know that we don't believe in the war," she said.

  “marched” and “rally” refer to the same demonstration event
Other Research Plans

Cross-document Information Processing and Reasoning
- Information Aggregation: gather together IE results from a set of relevant documents to produce compact tables ("time-aware")
- Information Correction: correct wrong information and recover missed information
- Information Prediction: predict events which are likely to happen in the future with probabilities
- Preliminary results on event extraction using IR feedback were promising (Ji and Grishman, ACL 2008)

Information-aware Machine Translation
- Develop co-ordinated name tagging of source and target bi-texts
- MT model from these name-tagged bitexts; information-driven decoding
My Main NLP Research Themes (1996-)

- Joint Inference for Natural Language Processing (this talk)
- Cross-lingual Information Extraction and Retrieval (joint work with UC Berkeley and Purdue; GALE work)
- Refine Event Extraction Using Cross-document Inference (post-doc work)
- Bootstrapping for Relation Extraction (post-doc work)
- Optimize Speech Segmentation for Information Extraction (joint work with UW and UC Berkeley; GALE work)

- Machine Learning
  - Semi-supervised Learning for Information Extraction
  - Cross-lingual Co-Training

- Semantic Role Labeling (intern work at IBM Research)
- Information Extraction for Financial Analysis (intern work at FactSet Research Systems)
- Text Summarization (Master thesis)
Thanks!