Monolingual Knowledge Acquisition and a Multilingual Information Environment

Kentaro Torisawa
Language Infrastructure Group
MASTAR Project
NICT, Japan
Self introduction

• Have been a group leader of the language infrastructure group, MASTAR project, NICT, Japan for two years.
  – [http://www2.nict.go.jp/x/x161/index-e.html](http://www2.nict.go.jp/x/x161/index-e.html)

• Research focus
  – Monolingual Knowledge Acquisition (KA) from the Web
  – development of applications using the acquired knowledge
Self introduction

• Have been a group leader of the language infrastructure group, MASTAR project, NICT, Japan for two years.
  – [http://www2.nict.go.jp/x/x161/index-e.html](http://www2.nict.go.jp/x/x161/index-e.html)

• Research focus
  – Monolingual Knowledge Acquisition (KA) from the Web
  – development of applications using the acquired knowledge

I have never done any research on MT...
Motivation behind This Talk

• Two fold

1. Monolingual Knowledge Acquisition (KA) reached the level such that interaction with the other fields may be fruitful

2. The methodologies in monolingual KA may give the MT community a new insight and novel applications
Contents

• The current status of our monolingual KA research
  – NICT Concept Dictionary
  – Application: recipe search system

• Possible interaction between MT and KA
  – Expansion of bilingual corpora
  – Bilingual Co-training
    • Monolingual KA method using translation
NICT Concept Dictionary

- Describing semantic relations between words
- Automatically acquired from the Web texts using automatic KA methods
- Currently 2.2 million Japanese words are covered
The Unknown
As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know.

D.H. Rumsfeld, Feb. 12, 2002,
Department of Defense news briefing
Looking for Trouble

• In early 2008, some people went to hospitals because of the fried dumplings polluted by pesticide

• A big media stir...
Polluted Fried Dumplings

- The concept dictionary could predict the food poisoning incident from polluted dumplings from the Web texts clawled before the incident.

“dumplings”

“Pesticide Residue” as possible trouble in eating dumplings
Less Serious Case

• If you want to go to Tokyo Disneyland...

“Height Restriction” for kids

“Disneyland”
Innovative Idea?

• If you want to lose weight....

• Tools for diet

“balloon”

“tomato”

“beer yeast”

“sugar”
How we constructed the concept dictionary

• Various knowledge acquisition methods
  – Hyponymy relation acquisition (Oh, et al., ACL 2009, Yamada et al., EMNLP 2009, Sumida, et al., LREC 2008)
  – EM-based word clustering (Kazama, et al., ACL 2008)
  – Generic Relation extraction (De Saeger et al., ICDM 2009, COLING 2008)
  – Verb entailment acquisition (Hashimoto, et al., EMNLP 2009)
  – Word class acquisition (De Saeger et al., IUCS 2009)
  – and more…
How we constructed the concept dictionary

• Various knowledge acquisition methods
  – Hyponymy relation acquisition (Oh, et al., ACL 2009, Yamada et al., EMNLP 2009, Sumida, et al., LREC 2008)
  – EM-based word clustering (Kazama, et al., ACL 2008)
  – Generic Relation extraction (De Saeger et al., ICDM 2009, COLING 2008)
  – Verb entailment acquisition (Hashimoto, et al., EMNLP 2009)
  – Word class acquisition (De Saeger et al., IUCS 2009)
  – and more...
Semantic Relation Acquisition

• A minimally supervised method for acquiring high-level semantic relations between noun pairs from the Web.

• Using this method we mined 50 million Japanese Web pages and obtained:
  • 30K causal relations with >80% precision (60K with >70%)
  • 30K product-material relations with >80% precision
  • 20K prevention relations with 74% precision
Outline of the Method

1. **Input:** a handful of *lexico-syntactic seed patterns* used to characterize the target relation
   - The system learns **reliable paraphrases** of the seed patterns using *class dependent pattern induction* (see later)

2. **Finally,** the system outputs a list of noun pairs ranked according to a confidence score

Outline of the Method:

- **Input seed patterns**
  - “avoiding X by Y”, “Y for preventing X”, ...

- **Extended patterns**
  - “anti-Y X”, “X that help prevent Y”, ...

- **Acquired relations**
  - “(ginseng, diabetes)”, “(laser, glaucoma)” ...

Automatically learn reliable paraphrases of the seed patterns

Extracting noun pairs matching the learnt patterns
Problems with the State-of-the-art

• Currently most state-of-the-art relation extraction systems (e.g. Espresso, Pantel et al. ‘06) perform pattern induction and instance extraction in a mutually recursive bootstrapping process, like so:

Input: **seed instances**
(tobacco, cancer),
(virus, influenza), ...

**Pattern extraction**: find linguistic contexts in which the seed instances co-occur: “X causes Y”, “X triggered Y”...

**Instance scoring** to filter out suspect instances

**Iterative Bootstrapping**

**Pattern induction**: learn reliable patterns that signal the target relation: “X is a source of Y”, ...

**Instance extraction**: use the induced patterns to extract new instances

Output: **acquired relation instances**
(drought, famine), ...

• We do not take such an approach since controlling the bootstrapping is extremely difficult
But Pattern Induction is Hard!

- Need to learn *high precision-high recall* patterns that co-occur with *a large number of correct* instances

**Google**

"caused"

- **Question**: How to deal with the ubiquitous, so-called "generic" patterns like "X by Y"?

  E.g. proper causal relation

  "death by drowning" ✔

  But also:  "new iPhone by Apple" ✗

  "registration by email" ✗

  "approval by committee" ✗

  "hotel by the sea" ✗
Class Dependent Pattern Induction

- **Class-dependent Patterns:** Make pattern induction class dependent by breaking patterns like “X by Y” up into word class dependent versions, i.e.

  - X by Y
    - [Class:Disease] by [Class:Chemical Substance]
    - [Class:Art Work] by [Class:Person]
    - [Class:Action] by [Class:Person]
    - [Class:Products] by [Class:Company]
    - ...

- **Key Assumption:** Class dependent patterns are not ambiguous.

- **Semantic word class information** can be obtained through large scale EM-based noun clustering (Kazama et al. ACL ’08)
Algorithm

Rank all the noun pairs observed in a single sentence according to the following score

$$Score(n_i, n_j, S) = \max classes(n_i), classes(n_j), \text{patterns } p \} \begin{align*}
CScore \times \text{Para} \times \text{Assoc}
\end{align*}$$

A noun pair \((n_i, n_j)\)’s final score is the best combination of the three component scores \(CScore, \text{Para} \) and \(\text{Assoc} \) maximized over:

- all \textbf{semantic classes} the noun pair belongs to, and
- all \textbf{class dependent patterns} the noun pair co-occurs with
Algorithm

Rank all the noun pairs observed in a single sentence according to the following score

$$Score(n_i, n_j, S) = \max \ classes(n_i), \ classes(n_j), \ \text{patterns} \ p \ \{ \frac{CScore \times Para \times Assoc}{\}$$

$CScore$ reflects how appropriate the semantic classes of $n_i$ and $n_j$ are for the target relation (a “class score”). This is measured by the overlap between the noun pairs that co-occur with the given seed patterns and all the possible combination of words in the class pair.
Algorithm

Rank all the noun pairs observed in a single sentence according to the following score

\[ \text{Score}(n_i, n_j, S) = \max \text{classes}(n_i), \text{classes}(n_j), \text{patterns } p \{ \]

\[ \quad C\text{Score} \times \text{Para} \times \text{Assoc} \]

\[ \} \]

*Para* scores pattern *p* as a *class dependent paraphrase* of the seed patterns. We regard two patterns as class dependent paraphrase if the two patterns have much overlap in the nouns that they co-occur inside particular noun classes.
Algorithm

Rank all the noun pairs observed in a single sentence according to the following score

\[ \text{Score}(n_i, n_j, S) = \max \text{classes}(n_i), \text{classes}(n_j), \text{patterns } p \{ \text{CScore} \times \text{Para} \times \text{Assoc} \} \]

**Assoc** measures the *association strength* between noun pairs and patterns (PMI)
Experiments: *Causal Relations*

- >80% precision @30K samples
- >70% precision @60K samples

Instances extracted by the seed patterns (2/3 raters)

Graph details:
- X-axis: Samples Ranked by Score
- Y-axis: Precision (%)
Experiments: *Causal Relations*

Some examples of acquired relations:

<table>
<thead>
<tr>
<th>class pair</th>
<th>rank</th>
<th>relation instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{471} \times c_{290}$</td>
<td>22</td>
<td>chiroshinaaze - sobakasu (tyrosinase - freckles)</td>
</tr>
<tr>
<td>$c_{468} \times c_{290}$</td>
<td>62</td>
<td>kabi - nioi (mold - bad smell)</td>
</tr>
<tr>
<td>$c_{468} \times c_{290}$</td>
<td>274</td>
<td>dani - hifu toraburu (mites - skin troubles)</td>
</tr>
<tr>
<td>$c_{471} \times c_{290}$</td>
<td>394</td>
<td>zanryu enso - kayumi (chlorine residue - itching)</td>
</tr>
<tr>
<td>$c_{475} \times c_{1}$</td>
<td>5889</td>
<td>nihonshu - himan (Japanese sake - obesity)</td>
</tr>
<tr>
<td>$c_{290} \times c_{290}$</td>
<td>6523</td>
<td>mushiba - koushuu (caries teeth - bad breath)</td>
</tr>
<tr>
<td>$c_{471} \times c_{1}$</td>
<td>17135</td>
<td>taurin - doumyaku kouka* (taurine - arterial sclerosis)</td>
</tr>
</tbody>
</table>
Real application: Recipe Search

• Beta service by NIFTY Co. (ISP)
  – Developed by a researcher and a programmer in three weeks

• Give advices in cooking with recipes

• Flexible search using semantic relations

http://labs.nifty.com/beta/recipe/
Recipe Search Using Concept Dictionary

• Collected 200,000 Recipes from Blog articles
  – Trained a classifier that judges if a given blog article describes a recipe
  – Basically, the classifier regards the article containing many ingredients as a recipe
  – More than 5,000 names of ingredients were extracted from the Concept Dictionary
Recipe Search Using Concept Dictionary

- Give advices using **many types of semantic relations in the concept dictionary**, which are extracted by the generic relation extraction system.

---

**Autumn** is the best season for saury.
How about a **jack** instead of saury? DHA contained in saury prevents lifestyle-related illness.
Query expansion using semantic relations

Query: "I’m suffering from Raynaud’s disease" (poor circulation of blood)

Semantic Relation in the Concept Dictionary
Garlic has a good effect on Raynaud’s disease

Provide a recipe document that includes garlic but Raynaud’s disease
So what do these things mean to MT research?

• Automatically generating/expanding bilingual corpora using paraphrase
  – Bond et al., IWSLT 2008, Nakov 2008, Mirkin et al., ACL-IJCNLP 2009

• The NICT Concept Dictionary contains knowledge supporting a wide range of paraphrases
  – Verb entailment (Manually checked. Hashimoto et al., EMNLP 2009)
    • チンする ⇒ 加熱する (microwave(verb) ⇒ heat(verb))
  – Paraphrase of class dependent patterns (De Saeger et al., ICDM 2009)
    • Y caused by X ⇔ Y by X

• Hyponymy relations
  • “Yatsuhashi” is a kind of “Japanese sweet”
  • “cream puff” is a kind of “sweet”
Paraphrasing bilingual corpora

• Verb entailment
  
  **Original corpus:** I heated the meal: 私は料理を温めた
  
  Expand using “温める⇒チンする” (heat ⇒ microwave)
  
  I heated the meal: 私は料理をチンした (I microwaved the meal)

• Class dependent patterns
  
  **Original corpus:** Cancer caused by asbestos: アスペストが引き起こした癌
  
  Expand using “Xが引き起こしたY = XによるY” (heat ⇒ microwave)
  
  Cancer caused by asbestos: アスペストによる癌

• Hyponymy relations
  
  **Original corpus:** I recommended a cream puff: シュークリームを薦めた
  
  Expand using “cream puff and Yatsuhashi are sweets.”
  
  I recommended a sweet named Yatsuhashi: ハツ橋を薦めた
Knowledge Gap?

• There is often a “multilingual knowledge gap” between countries (or languages)
• But the gap is often unknown unknown

Translated Queries

170,000 documents >> 8,400 documents

Recognition of multilingual unknown unknowns may boost needs for translation!
Multilingual Concept Dictionary

• By constructing multilingual concept dictionary, in which words in different languages are linked with each other, using MT technologies,
• we may be able to obtain a more complete knowledge base that can suggest multilingual unknown unknowns, and
• needs for machine translation may become more apparent
Mul2lingual Knowledge Acquisi2on?

• KA may benefit from MT systems
  – Monolingual KA processes can be applied to translated text archives
  – Monolingual KA can be integrated with translation

⇒ First Attempt: Bilingual Co-training
  – Translation was done by dictionary look-up
  – Oh et al., ACL-IJCNLP 2009
Bilingual Co-training

• Sample Task
  – Hyponymy relation Acquisition from the Wikipedia
    • When A is a kind of B, A and B are said to have a hyponymy relation
    • Example
      – (Tiger, Siberian Tiger)
      – (Country, Japan)
      – (Car Manufacturer, Toyota)
  • Supervised classification task
  • Use a layout structure in the Wikipedia as features for the classification
Hyponymy Relation Acquisition from Wikipedia

• From hierarchical layout structure of Wikipedia articles (Sumida et al, LREC 2008)
  – 39 M English candidates and 10 M Japanese ones

<table>
<thead>
<tr>
<th>Tiger</th>
<th>Range</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiger</td>
<td>Range</td>
<td>Taxonomy</td>
</tr>
<tr>
<td>Range</td>
<td>Taxonomy</td>
<td></td>
</tr>
<tr>
<td>Taxonomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subspecies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ Bengal tiger</td>
<td>▪ Malayan tiger</td>
<td>▪ Siberian tiger</td>
</tr>
</tbody>
</table>

Hyponymy relation candidates:
- (Tiger, Range)
- (Tiger, Taxonomy)
- (Tiger, Bengal tiger)
- (Tiger, Malayan tiger)
- (Tiger, Siberian tiger)

... (Subspecies, Siberian tiger)
Hyponymy Relation Acquisition from Wikipedia

• Binary classification of hyponymy-relation candidates
  – (hyper, hypo) → “Hyponymy relation” or “not”
    • (Tiger, Siberian tiger) → “hyponymy relation”
    • (Tiger, Taxonomy) → “not hyponymy relation”

• SVMs as classifiers (Sumida, 2008)
  – Lexical features
  – Structure-based features
    • Layout structure
    • Tree structure (our proposed one)
Basic Idea

- Is the classification accuracy improved by adding to the training data the data translated from another language?

- Maybe, translation of the reliable classification results can be used as well

- And the other direction may work as well...
Concept of Bilingual co-training

Language 1

- Further Enlarged Training Data for Language 1
- Classifier
- Enlarged Training Data for Language 1
- Manually Prepared Training Data for Language 1

Language 2

- Further Enlarged Training Data for Language 2
- Classifier
- Enlarged Training Data for Language 2
- Manually Prepared Training Data for Language 2

Translate reliable parts of classification results

Iteration

ACL-IJCNLP 09
Data

- Randomly selected 24,000 candidates
  - manually tagged as “positive” and “negative”
  - “positive sample”: “negative sample” = 1:2

<table>
<thead>
<tr>
<th>Set</th>
<th>En</th>
<th>Ja</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>20,000</td>
<td>20,000</td>
<td>For the initial classifier</td>
</tr>
<tr>
<td>Development</td>
<td>2,000</td>
<td>2,000</td>
<td>For optimal parameters</td>
</tr>
<tr>
<td>Blind test</td>
<td>2,000</td>
<td>2,000</td>
<td>Evaluating systems</td>
</tr>
</tbody>
</table>
“Is bilingual co-training better than a monolingual method?”

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th></th>
<th>Japanese</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYT</td>
<td>78.5</td>
<td>63.8</td>
<td>70.4</td>
<td>75.0</td>
<td>77.4</td>
<td>76.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INIT</td>
<td>77.9</td>
<td>67.4</td>
<td>72.2</td>
<td>74.5</td>
<td>78.5</td>
<td>76.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAN</td>
<td>76.8</td>
<td>70.3</td>
<td>73.4</td>
<td>76.7</td>
<td>79.3</td>
<td>78.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BICO</td>
<td>78.0</td>
<td>83.7</td>
<td>80.7</td>
<td>78.3</td>
<td>85.2</td>
<td>81.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- SYT: our implementation of (Sumida, 2008)
- INIT: our monolingual initial classifier
- BICO: classifier based on bilingual co-training
- BICO outperforms both SYT and INIT
  - 5.0—10.3% in $F_1$
“Is bilingual co-training better than simply translating training data?”

<table>
<thead>
<tr>
<th></th>
<th><strong>English</strong></th>
<th></th>
<th><strong>Japanese</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
<td>$P$</td>
</tr>
<tr>
<td>SYT</td>
<td>78.5</td>
<td>63.8</td>
<td>70.4</td>
<td>75.0</td>
</tr>
<tr>
<td>INIT</td>
<td>77.9</td>
<td>67.4</td>
<td>72.2</td>
<td>74.5</td>
</tr>
<tr>
<td>TRAN</td>
<td>76.8</td>
<td>70.3</td>
<td>73.4</td>
<td>76.7</td>
</tr>
<tr>
<td>BICO</td>
<td>78.0</td>
<td>83.7</td>
<td>80.7</td>
<td>78.3</td>
</tr>
</tbody>
</table>

- TRAN (English: 20,729 and Japanese: 20,486)
  - Translating training data in language S to language T
  - Adding the translation as newly labeled data to language T

- BICO outperforms TRAN
  - 3.6—7.3% in $F_1$
Bilingual Co-training: Summary

• Translation and learning are tightly coupled
  – Better than simple translation of training data
  – Though the translation is a simple look-up of bilingual dictionary

• Probably applicable to many classification tasks

• Future work: Integrating full MT systems in KA (in a non-trivial way)
Conclusion

- Our Knowledge Acquisition Methods from the Web
  - NICT Concept Dictionary
- Applications: Recipe Search
- Possible Interaction with MT researches
  - Bilingual Co-training
  - Multilingual concept dictionary for finding multilingual unknown unknowns?