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
- 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
- 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

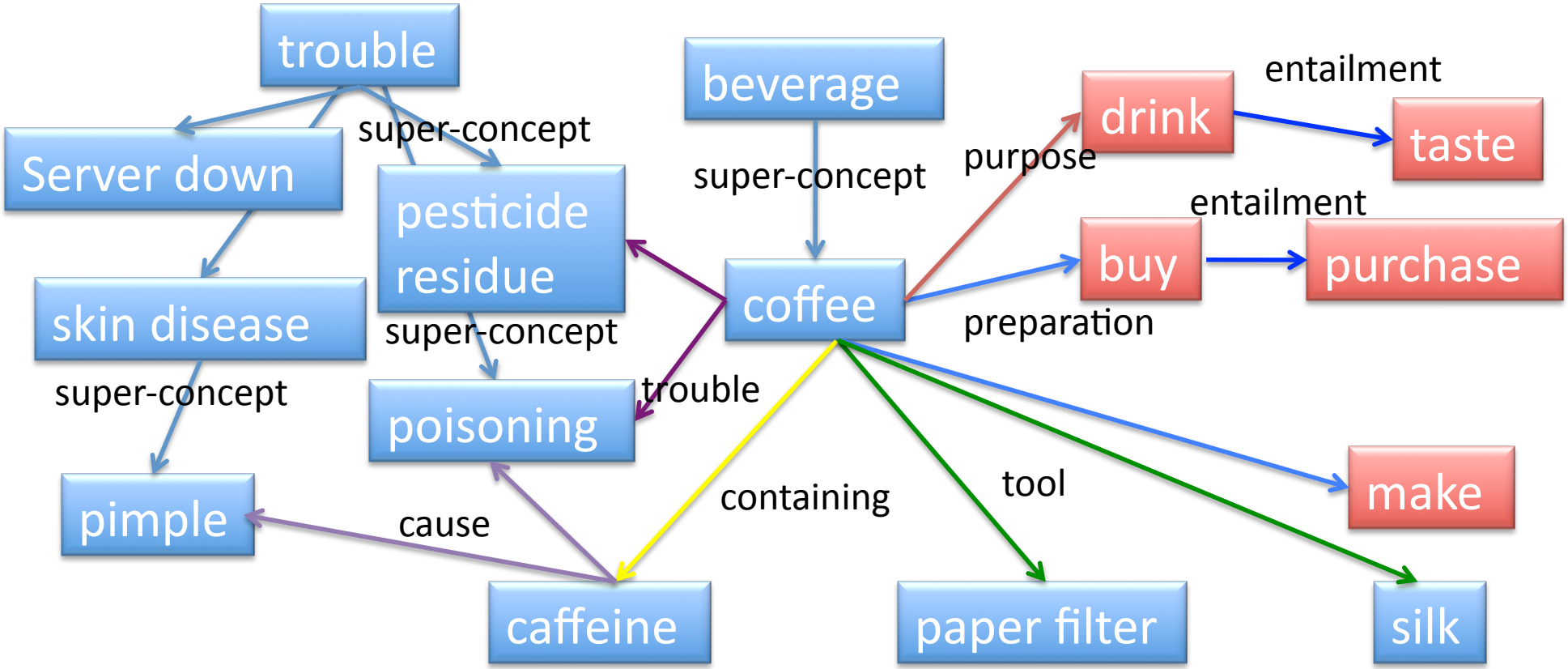
- Two fold
  1. Monolingual Knowledge Acquisition (KA) reached the level such that interaction with the other fields may be fruitful
  2. The methodologies in monolingual KA may give the MT community a new insight and novel applications

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



NICT Concept Dictionary: Example

- Support to find valuable “unknown unknowns” from the Web

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

Try YouTube

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

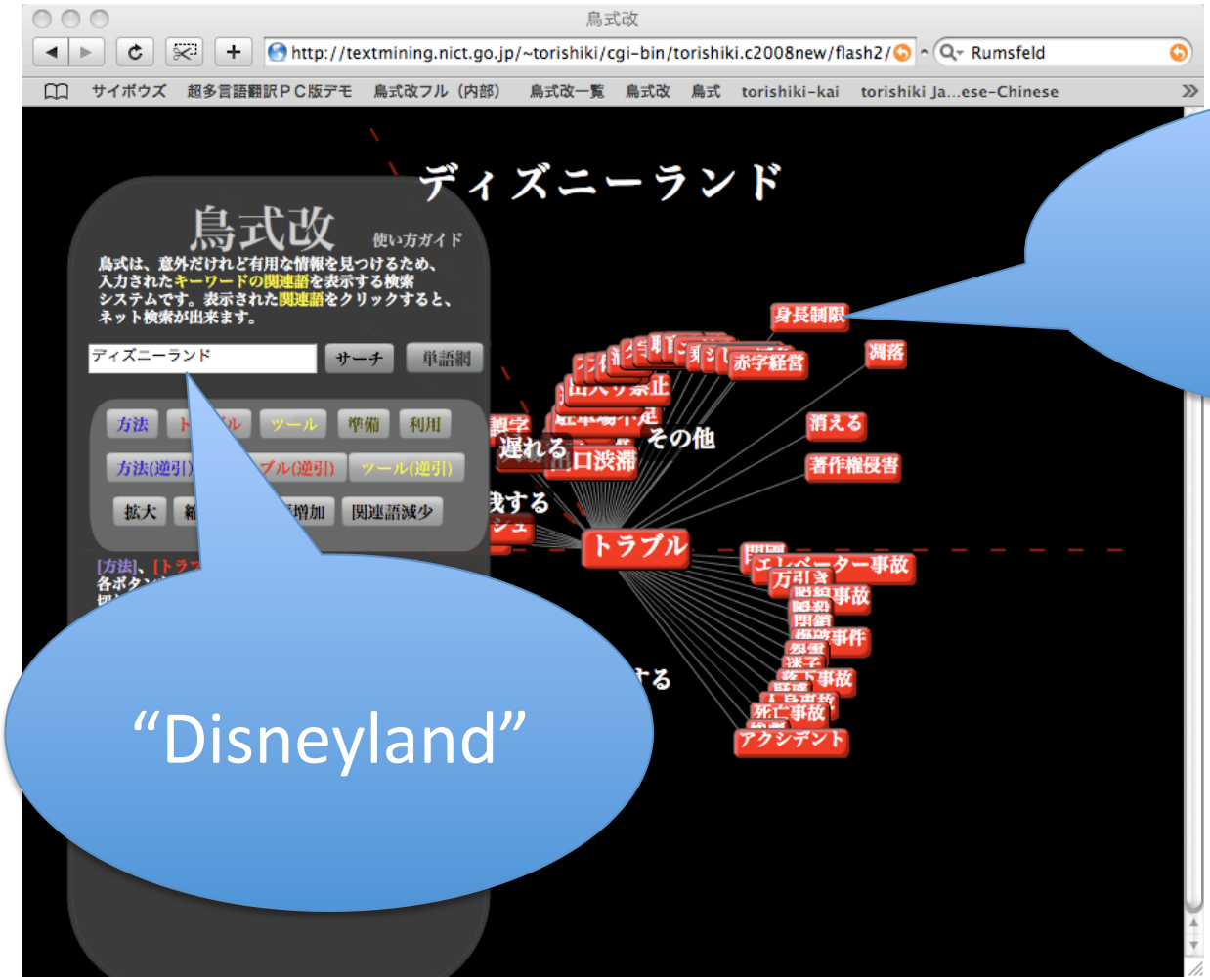
- The concept dictionary could predict the food poisoning incident from polluted dumplings from the Web texts clawled before the incident

"dumplings"

"Pesticide Residue" as possible trouble in eating dumplings

Less Serious Case

- If you want to go to Tokyo Disneyland…

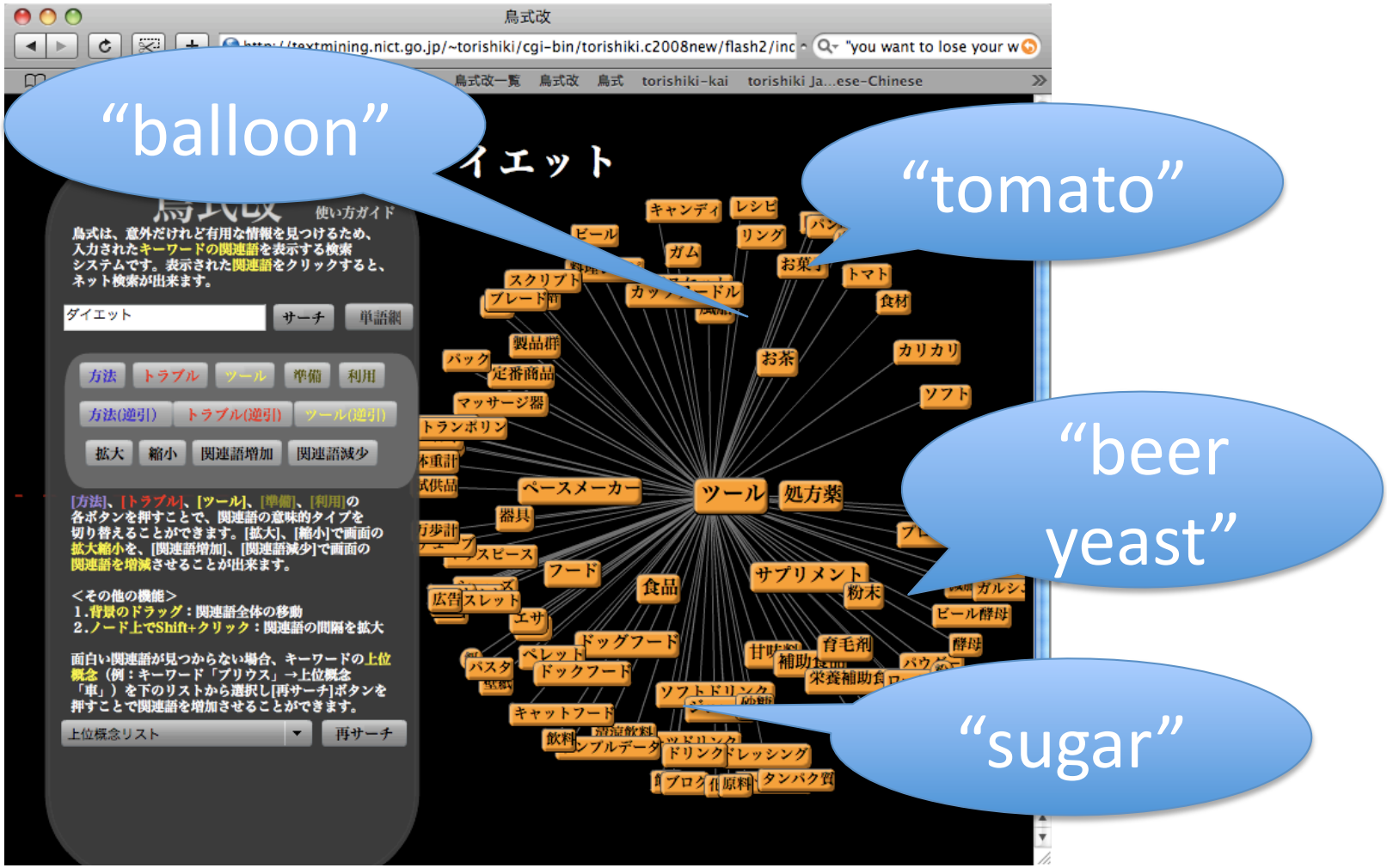


"Height Restriction" for kids

"Disneyland"

Innovative Idea？

- If you want to lose weight….
- Tools for diet



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

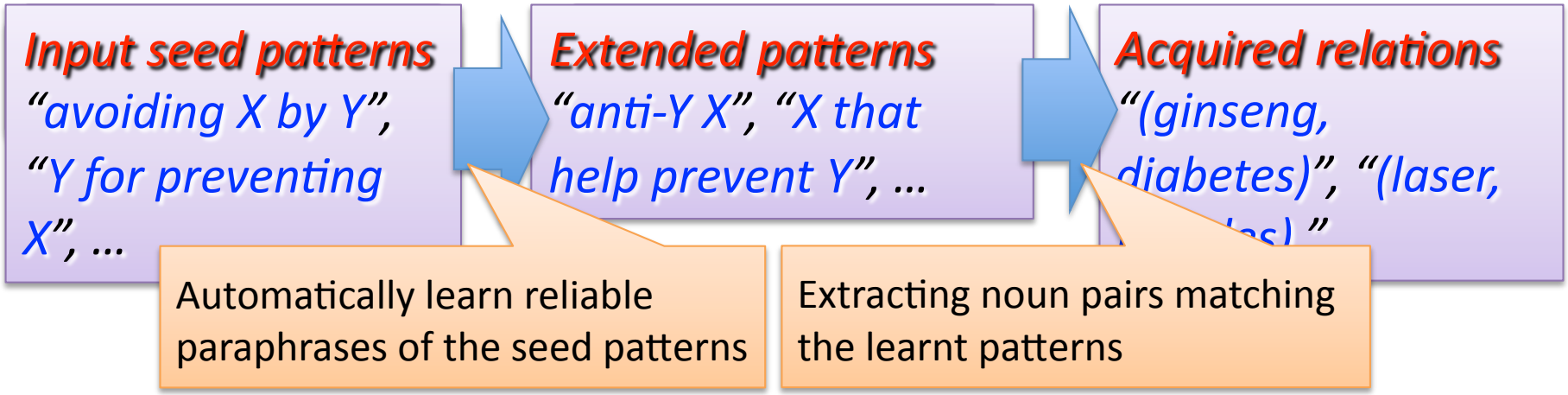
- A minimally supervised method for acquiring high-level semantic relations between noun pairs from the Web.
- Using this method we mined 50 million Japanese Web pages and obtained:
  - 30K causal relations with >80% precision (60K with >70%)
  - 30K product-material relations with >80% precision
  - 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

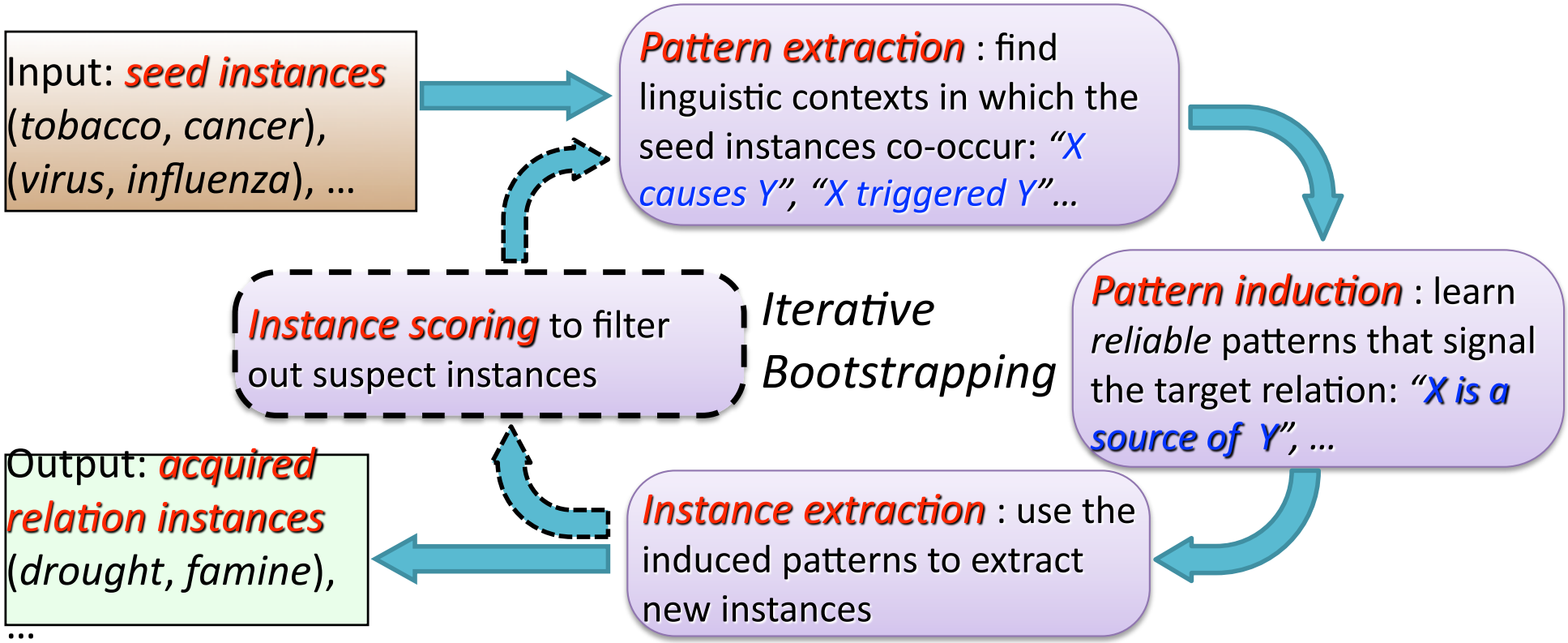
   •The system learns reliable paraphrases of the seed patterns using *class dependent pattern induction* (see later)

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



Problems with the State-of-the-art

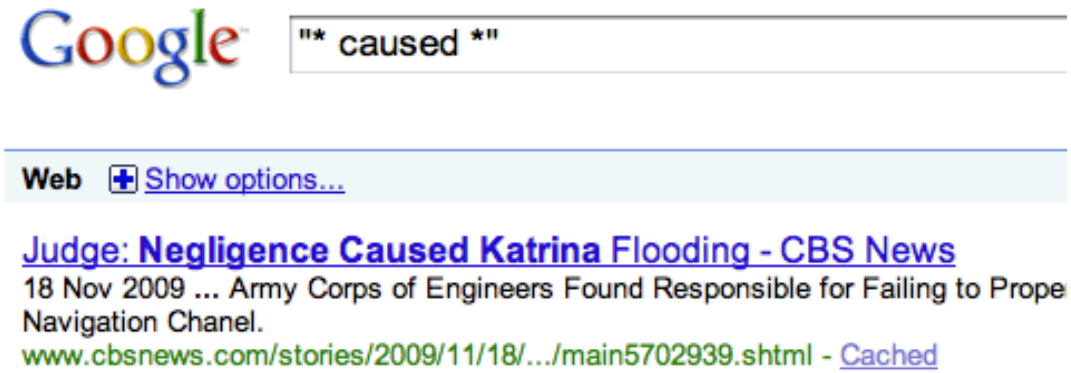
• Currently most state-of-the-art relation extraction systems (e.g. *Espresso*, Pantel et al. '06) perform pattern induction and instance extraction in a mutually recursive bootstrapping process, like so:



• 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



- 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" ✘

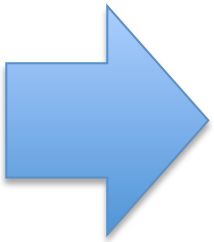
"approval by committee" ✘

"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

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

A noun pair (n_i, n_j)'s final score is the best combination of the three component scores *CScore*, *Para* and *Assoc* maximized over:

- all semantic classes the noun pair belongs to, and
- all 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

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

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

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

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.

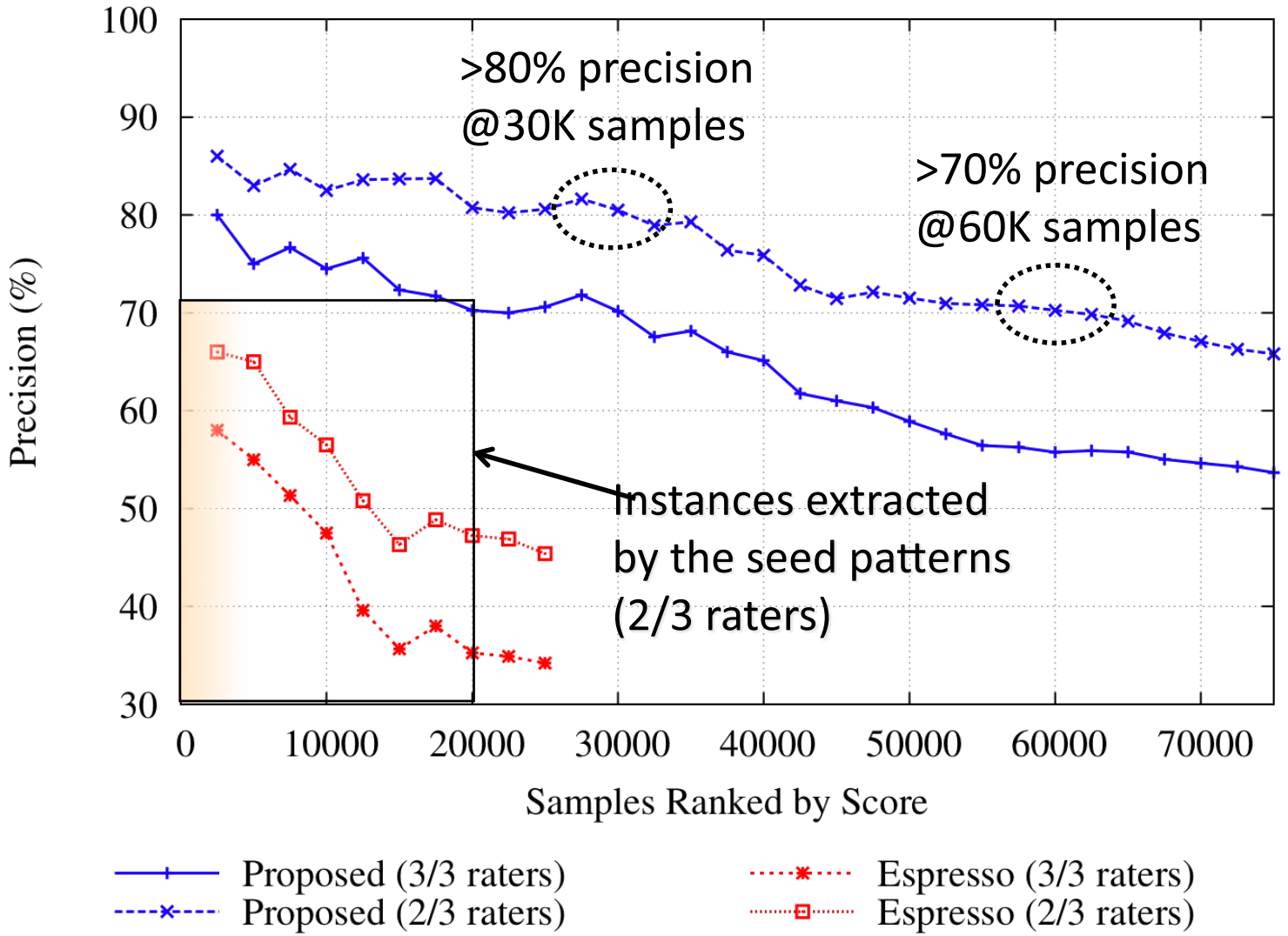
Algorithm

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

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

Assoc measures the *association strength* between noun pairs and patterns (*PMI*)

Experiments: *Causal Relations*



Experiments: *Causal Relations*

Some examples of acquired relations:

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

Real application: Recipe Search

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

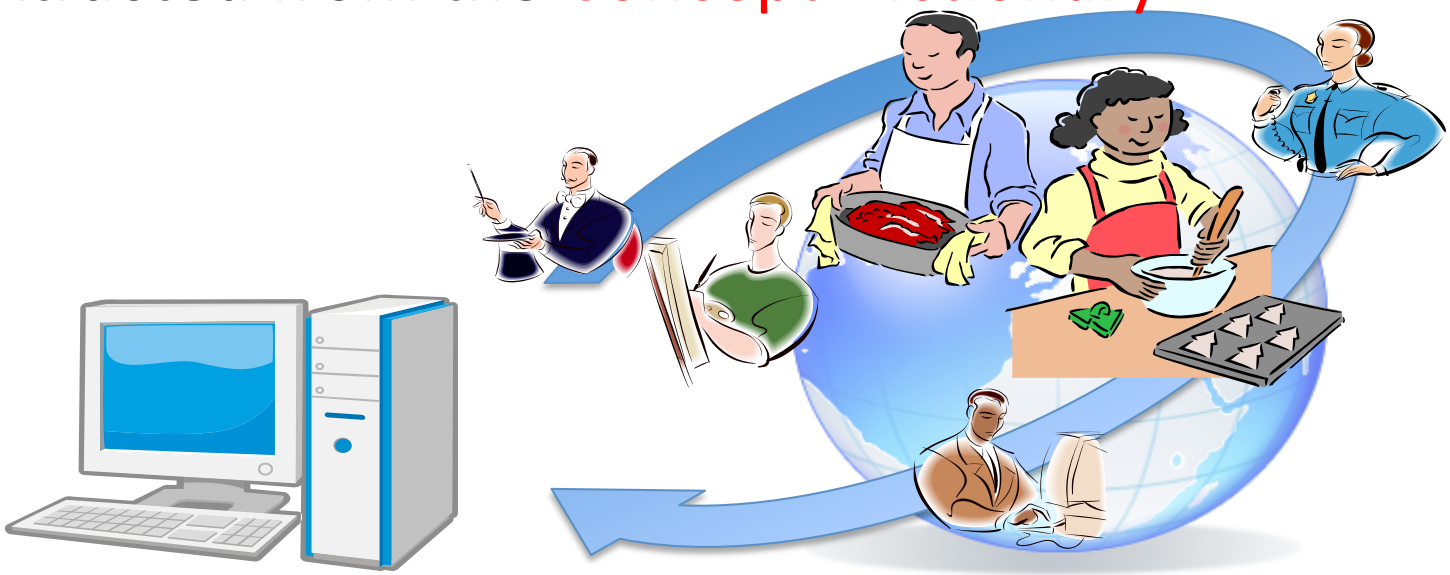
- Give advices in cooking with recipes

- Flexible search using semantic relations

http://labs.nifty.com/beta/recipe/

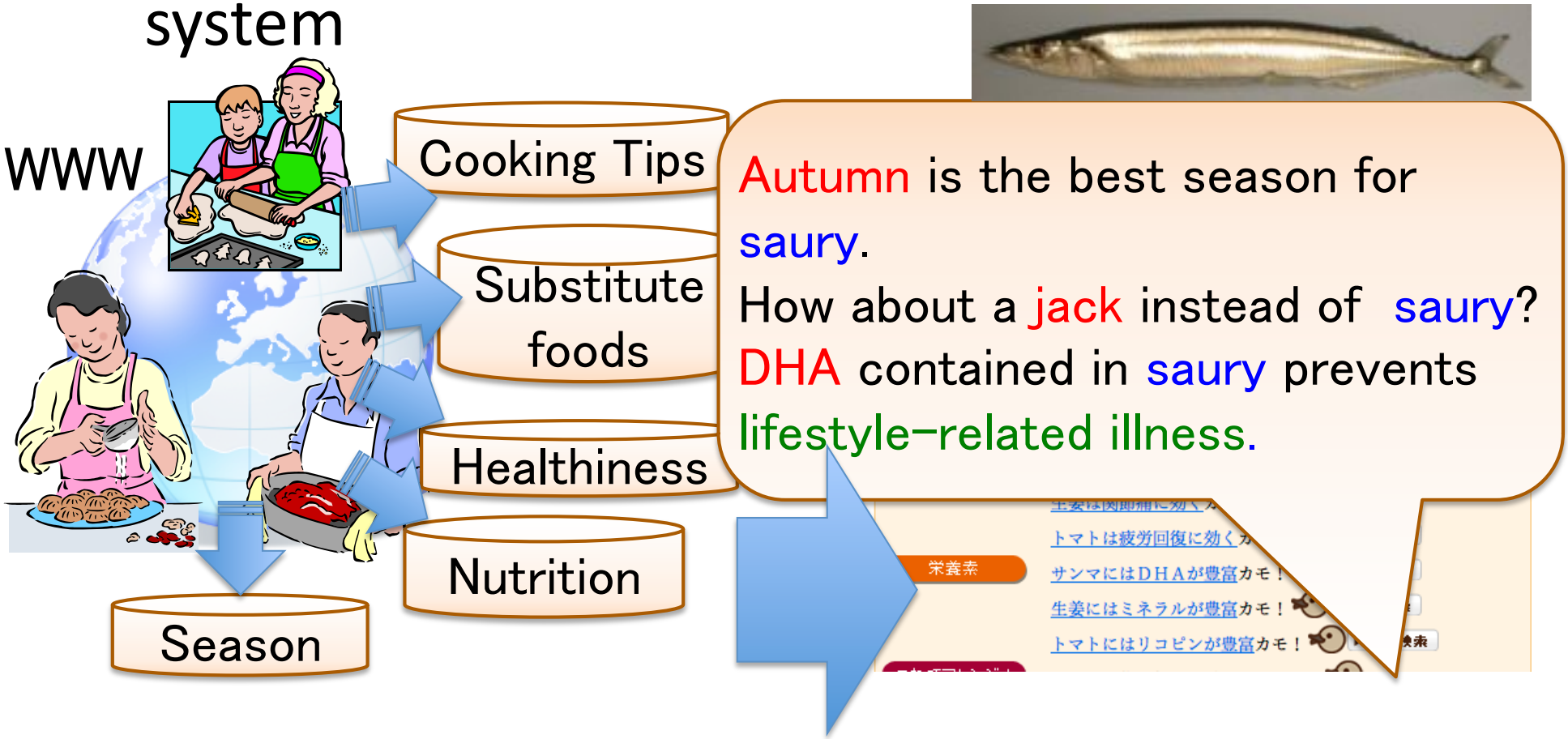
Recipe Search Using Concept Dictionary

- 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



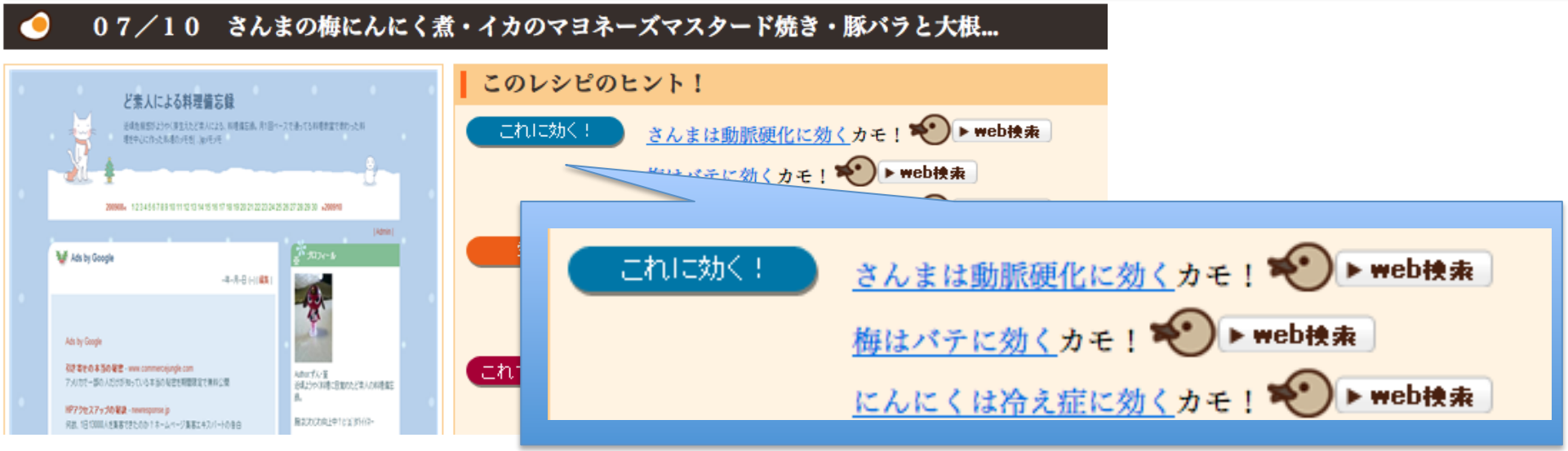
Query expansion using semantic relations

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 diseaese



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 : アスベストが引き起こした癌

Expand using “XがY引き起こした = XによるY” (heat ⇒ microwave)

Cancer caused by asbestos : アスベストによる癌

- Hyponymy relations

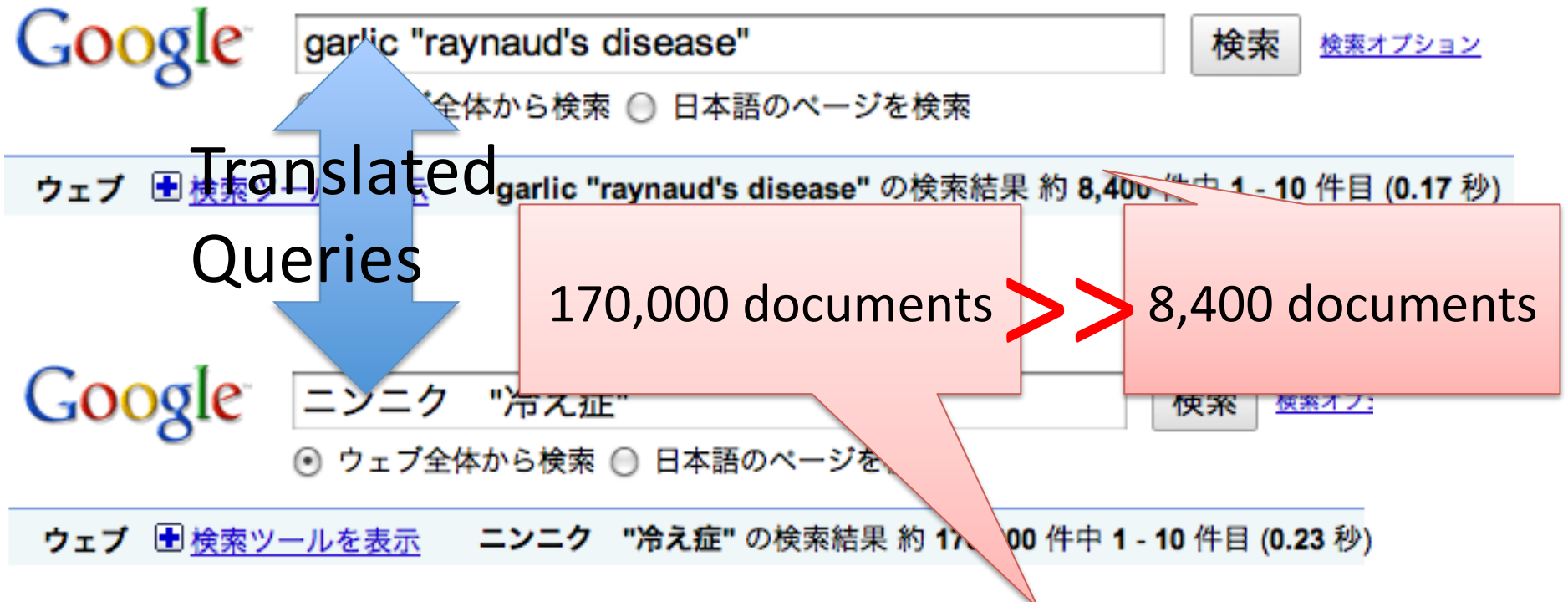
Original corpus: I recommended a cream puff : シュークリームを薦めた

Expand using “cream puff and *Yatsuhashi* are sweets.”

I recommended a sweet named *Yatuhashi* : 八ツ橋を薦めた

Knowledge Gap?

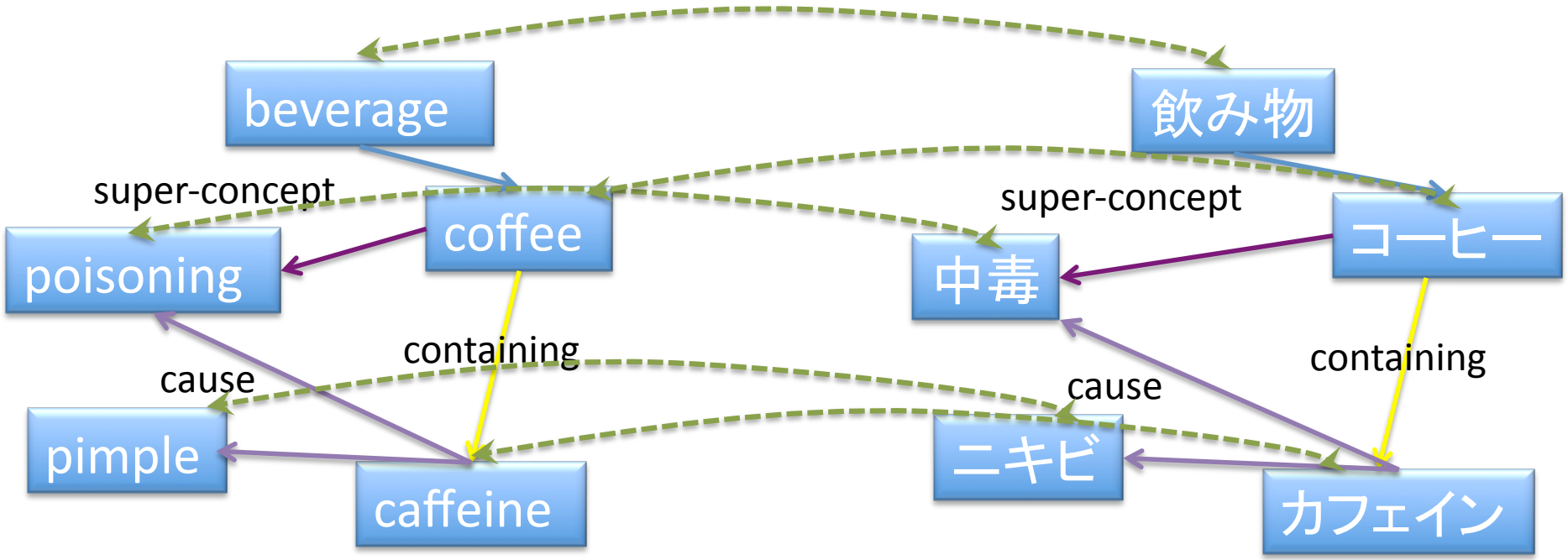
- There is often a “multlingual knowledge gap” between countries (or languages)
- But the gap is often unknown unknown



Recognition of multilingual unknown unknowns may boost needs for translation!

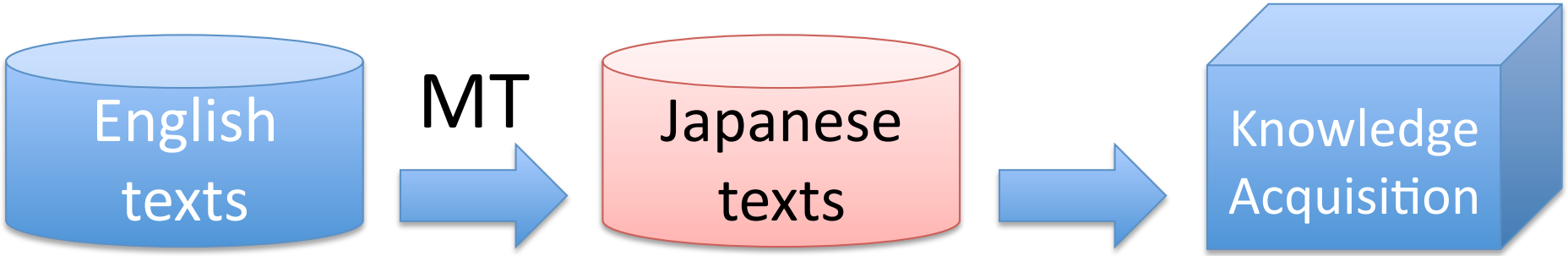
Multilingual Concept Dictionary

- 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



Multilingual Knowledge Acquisition?

- 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

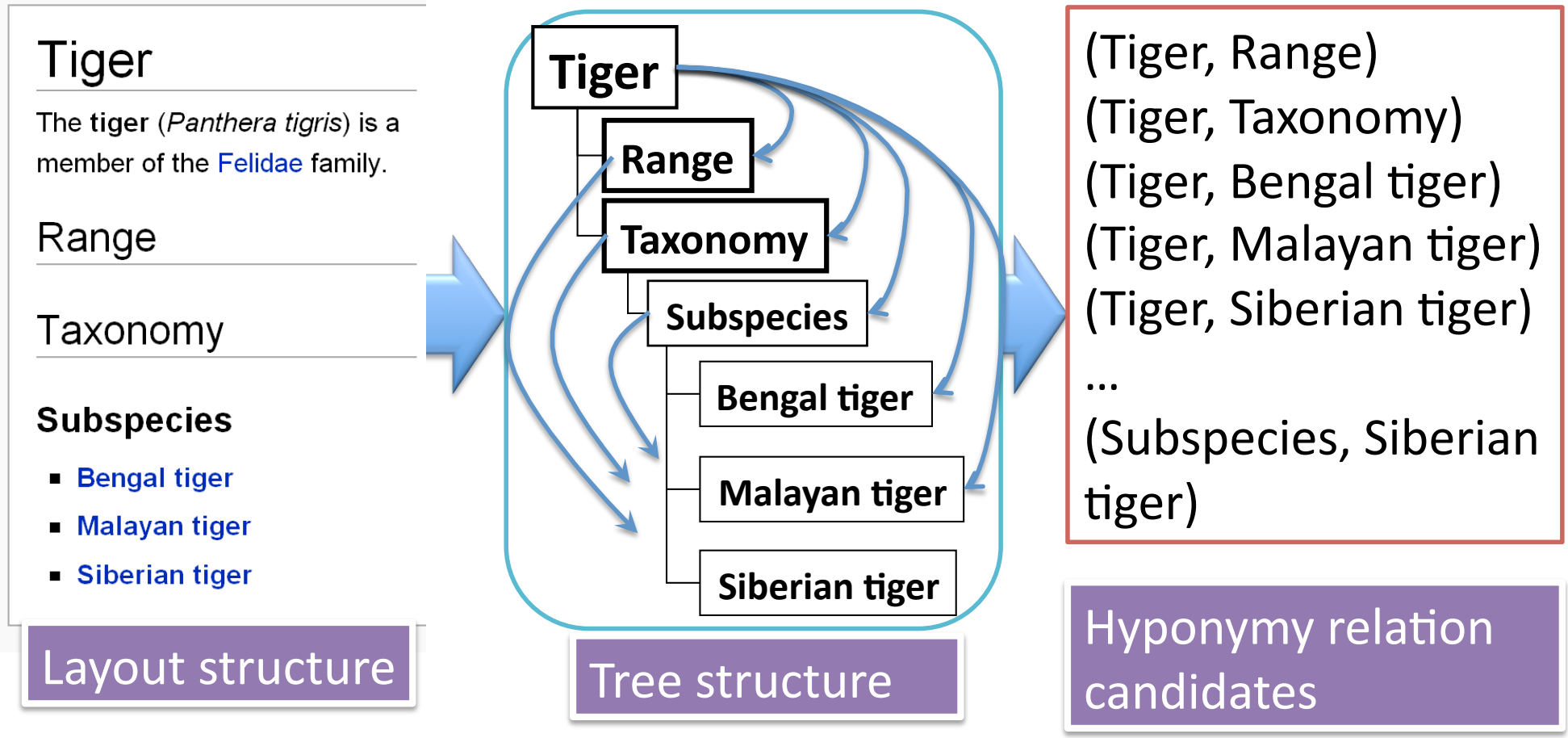
- Translation was done by dictionary look-up
- Oh et al., ACL-IJCNLP 2009

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

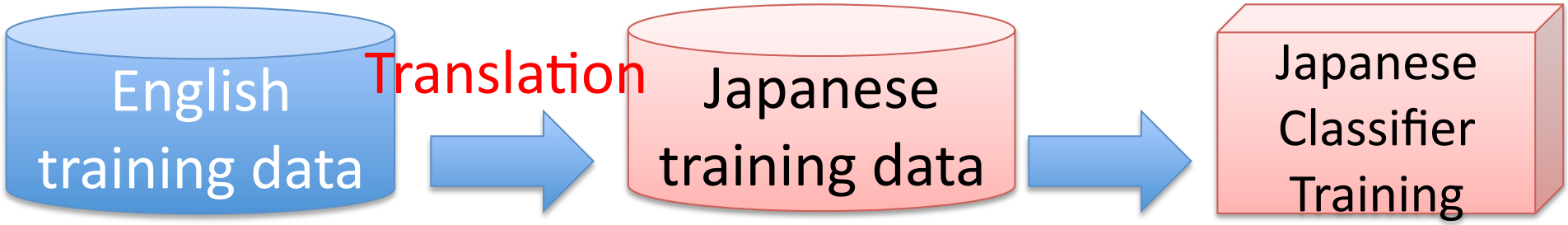


Hyponymy Relation Acquisition from Wikipedia

- 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

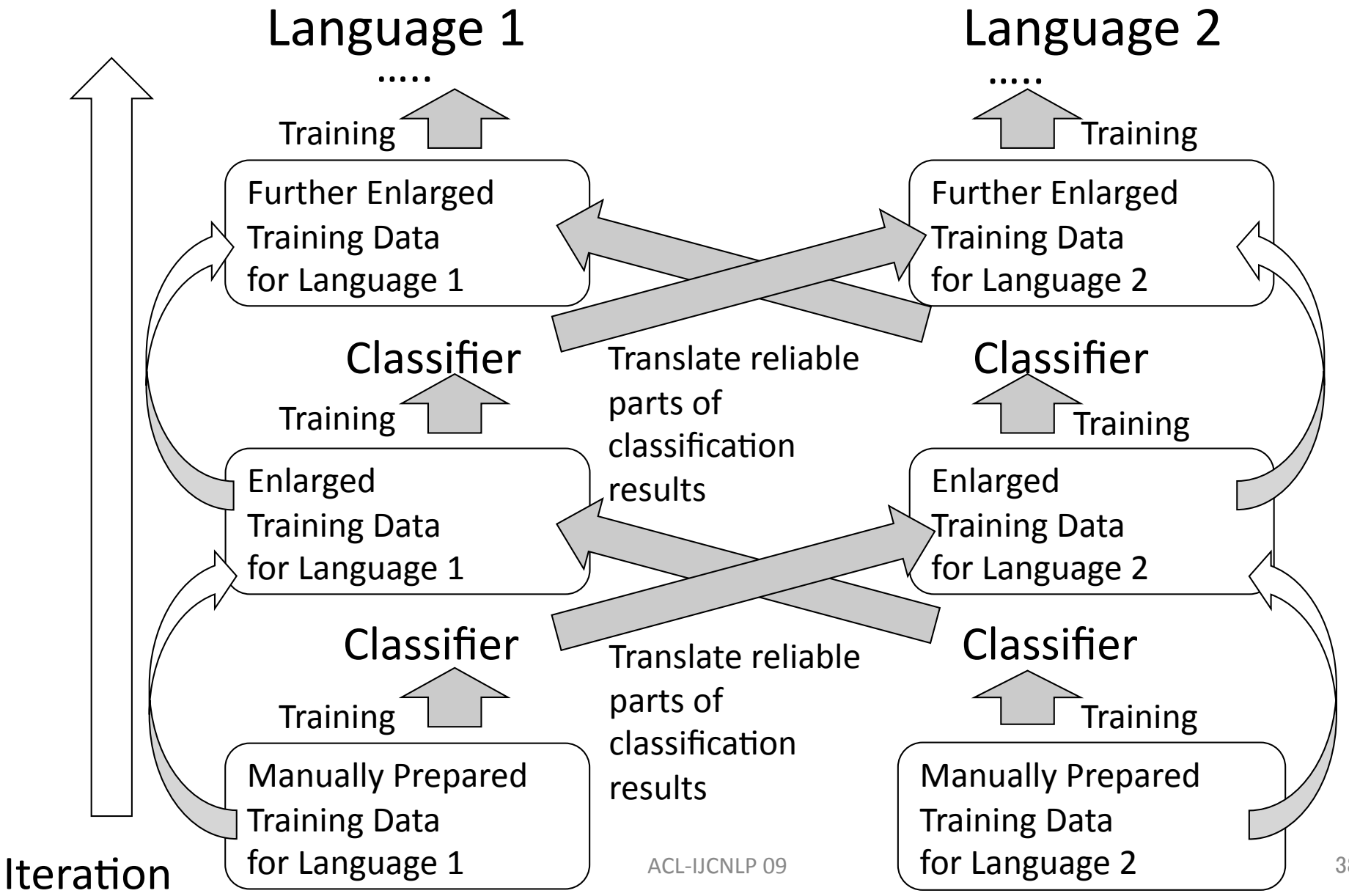
- Is the classification accuracy improved by adding to the training data the data translated from another language?



- 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



Data

- May, 2008 version of English Wikipedia
- June, 2008 version of Japanese Wikipedia
- Randomly selected 24,000 candidates
  - manually tagged as “positive” and “negative”
    - “positive sample”: “negative sample” = 1:2

Set	En	Ja	Purpose
Train	20,000	20,000	For the initial classifier
Development	2,000	2,000	For optimal parameters
Blind test	2,000	2,000	Evaluating systems

"Is *bilingual co-training* better than a monolingual method?"

	ENGLISH			JAPANESE		
	P	R	F_1	P	R	F_1
SYT	78.5	63.8	70.4	75.0	77.4	76.1
INIT	77.9	67.4	72.2	74.5	78.5	76.6
TRAN	76.8	70.3	73.4	76.7	79.3	78.0
BICO	78.0	83.7	80.7	78.3	85.2	81.6

- 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?"

	ENGLISH			JAPANESE		
	P	R	F_1	P	R	F_1
SYT	78.5	63.8	70.4	75.0	77.4	76.1
INIT	77.9	67.4	72.2	74.5	78.5	76.6
TRAN	76.8	70.3	73.4	76.7	79.3	78.0
BICO	78.0	83.7	80.7	78.3	85.2	81.6

- 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
  - Mutlilingual concept dictionary for finding multilingual unknown unkonwns?