German-Japanese Workshop on NLP for Information Management and Semantic Web (July 4-5, 2003)

Word Sense Acquisition from Bilingual Comparable Corpora

Hiroyuki Kaji
Central Research Laboratory, Hitachi, Ltd.

This research was supported by the New Energy and Industrial Technology Development Organization of Japan.

Overview of the talk

1. Introduction
   Motivation, goal and related work
2. Unsupervised word sense disambiguation using bilingual comparable corpora
3. Proposed method for clustering translation equivalents of a polysemous word
4. Experiment using Wall Street Journal and *Nihon Keizai Shimbun* corpora
5. Discussion
   Advantages and limitations
6. Conclusion

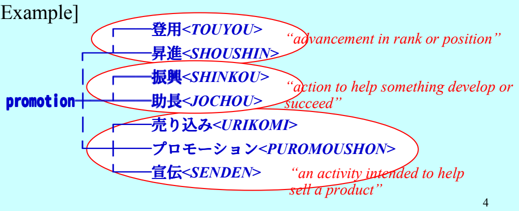
1.1 Motivation

- ■ Word sense disambiguation
  - – A subtask necessary for most NLP tasks, esp. MT and IR
  - – Great deal of research has been done over the past decade.
- ■ Word sense acquisition
  - – Human activity
  - – Inventories of word senses have been constructed by lexicographers based on their intuition.
  - – Problems with manual construction
    - High cost
    - Arbitrary division of word senses
    - Mismatch to application domains

1.2 Goal

- ■ Unsupervised word sense disambiguation using bilingual comparable corpora (Kaji and Morimoto, COLING2002)
- ■ Automatic word sense acquisition
  - – Cluster translation equivalents of a polysemous word to divide and define the senses of that word.

[Example]


promotion
登用<TOUYOU>
昇進<SHOUSHIN>
振興<SHINKOU>
助長<JOCHOU>
売り込み<URIKOMI>
プロモーション<PUROMOUSHON>
宣伝<SENDEN>
“advancement in rank or position”
“action to help something develop or succeed”
“an activity intended to help sell a product”

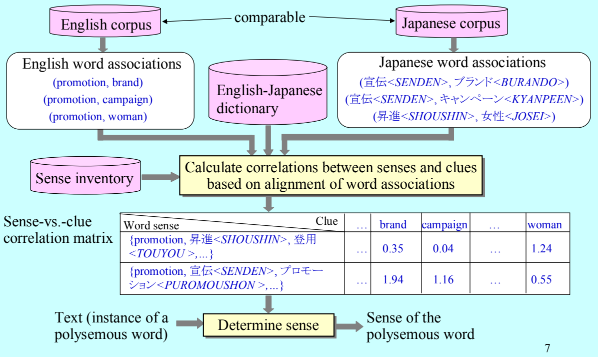
1.3 Related work

- ■ Overlapping distributional word clustering—define word senses using sets of synonyms
  - ● Fukumoto and Tsujii, COLING 1994
    - — Cluster synonyms of each target polysemous verb.
    - — Each cluster represents a sense of the target word.
  - ● Pantel and Lin, KDD 2002
    - — Cluster all nouns with occurrence frequencies larger than a threshold.
    - — A polysemous word is assigned to multiple clusters, each of which represents one of its senses.
- ■ Word sense discrimination
  - ● Schuetze, Computational Linguistics 1998
    - — Cluster documents containing each target polysemous word.
    - — Each document cluster corresponds to a sense of the target word. However it is not labeled.

Overview of the talk

1. Introduction
   Motivation, goal and related work
2. Unsupervised word sense disambiguation using bilingual comparable corpora
3. Proposed method for clustering translation equivalents of a polysemous word
4. Experiment using Wall Street Journal and *Nihon Keizai Shimbun* corpora
5. Discussion
   Advantages and limitations
6. Conclusion

Unsupervised WSD using bilingual comparable corpora



English corpus ← comparable → Japanese corpus

English word associations
(promotion, brand)
(promotion, campaign)
(promotion, woman)

English-Japanese dictionary

Japanese word associations
(宣伝<*SENDEN*>, ブランド<*BURANDO*>)
(宣伝<*SENDEN*>, キャンペーン<*KYANPEEN*>)
(昇進<*SHOUSHIN*>, 女性<*JOSEI*>)

Sense inventory

Calculate correlations between senses and clues based on alignment of word associations

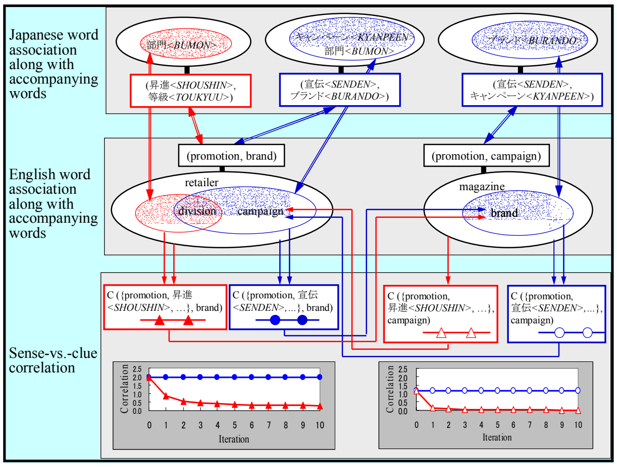
Sense-vs.-clue correlation matrix

Word sense \ Clue	…	brand	campaign	…	woman
{promotion, 昇進<*SHOUSHIN*>, 登用<*TOUYOU* >,…}	…	0.35	0.04	…	1.24
{promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON* >,…}	…	1.94	1.16	…	0.55

Text (instance of a polysemous word) → Determine sense → Sense of the polysemous word

Iterative calculation of sense-vs.-clue correlations

- Problems in WSD using bilingual comparable corpora
  - Ambiguity in alignment of word associations
  - Failure in alignment of word associations caused by
    - Disparity of topical coverage between the texts of different languages
    - Incomplete coverage of the bilingual dictionary
- Solution
  - Define correlation between a sense of the target word and a clue as the mutual information of the target word and the clue multiplied by the maximum plausibility of alignments suggesting the sense-clue pair.
  - Calculate the correlations iteratively based on
    - Assumption 1: Plausibility of an alignment of word associations depends on plausibility of alignments between words accompanying those word associations.
    - Assumption 2: The correlation between a sense and a clue depends on the correlations between that sense and clues accompanying that clue.



Japanese word association along with accompanying words

English word association along with accompanying words

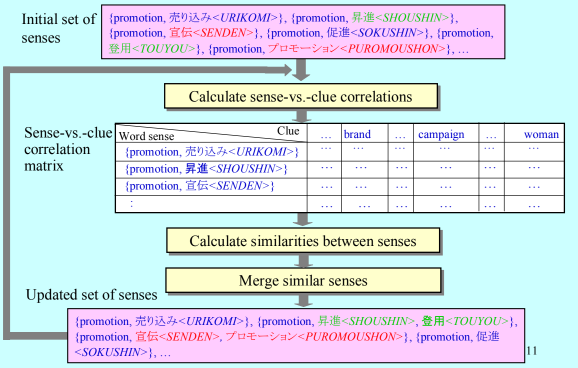
Sense-vs.-clue correlation

部門<*BUMON*>
キャンペーン<*KYANPEEN*>
部門<*BUMON*>
ブランド<*BURANDO*>
(昇進<*SHOUSHIN*>, 等級<*TOUKYUU*>)
(宣伝<*SENDEN*>, ブランド<*BURANDO*>)
(宣伝<*SENDEN*>, キャンペーン<*KYANPEEN*>)
(promotion, brand)
(promotion, campaign)
retailer
division
campaign
magazine
brand
C ({promotion, 昇進<*SHOUSHIN*>, …}, brand)
C ({promotion, 宣伝<*SENDEN*>,…}, brand)
C ({promotion, 昇進<*SHOUSHIN*>, …}, campaign)
C ({promotion, 宣伝<*SENDEN*>,…}, campaign)
Correlation
2.5
2.0
1.5
1.0
0.5
0.0
0 1 2 3 4 5 6 7 8 9 10
Iteration
Correlation
2.5
2.0
1.5
1.0
0.5
0.0
0 1 2 3 4 5 6 7 8 9 10
Iteration

Overview of the talk

1. Introduction
   Motivation, goal and related work
2. Unsupervised word sense disambiguation using bilingual comparable corpora
3. Proposed method for clustering translation equivalents of a polysemous word
4. Experiment using Wall Street Journal and *Nihon Keizai Shimbun* corpora
5. Discussion
   Advantages and limitations
6. Conclusion

Translingual distributional clustering of translation equivalents



Initial set of senses

{promotion, 売り込み<*URIKOMI*>}, {promotion, 昇進<*SHOUSHIN*>}, {promotion, 宣伝<*SENDEN*>}, {promotion, 促進<*SOKUSHIN*>}, {promotion, 登用<*TOUYOU*>}, {promotion, プロモーション<*PUROMOUSHON*>}, …

Calculate sense-vs.-clue correlations

Sense-vs.-clue correlation matrix

Word sense \ Clue	…	brand	…	campaign	…	woman
{promotion, 売り込み<*URIKOMI*>}	…	…	…	…	…	…
{promotion, **昇進**<*SHOUSHIN*>}	…	…	…	…	…	…
{promotion, 宣伝<*SENDEN*>}	…	…	…	…	…	…
:	…	…	…	…	…	…

Calculate similarities between senses

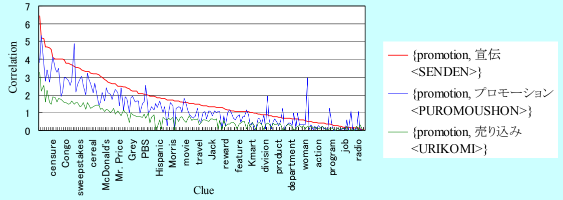
Merge similar senses

Updated set of senses

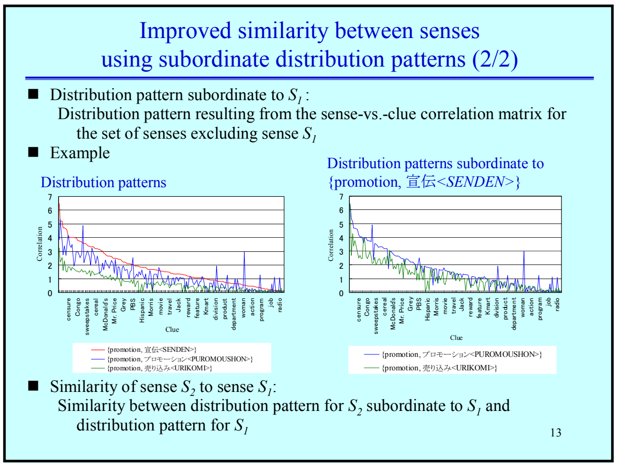
{promotion, 売り込み<*URIKOMI*>}, {promotion, 昇進<*SHOUSHIN*>, **登用**<*TOUYOU*>}, {promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>}, {promotion, 促進<*SOKUSHIN*>}, …

Improved similarity between senses using subordinate distribution patterns (1/2)

- Weakness of translingual distribution patterns:
  A clue has always high correlation with only one sense, and therefore translation equivalents representing the same sense do not necessarily have very similar patterns.
- Example



  - All these senses define the “sales activity” sense of “promotion.”
  - Most clues for identifying that sense have the highest correlations with {promotion, **宣伝**<*SENDEN*>}, which is the most dominant translation equivalent of “promotion” in the corpus.



Improved similarity between senses using subordinate distribution patterns (2/2)

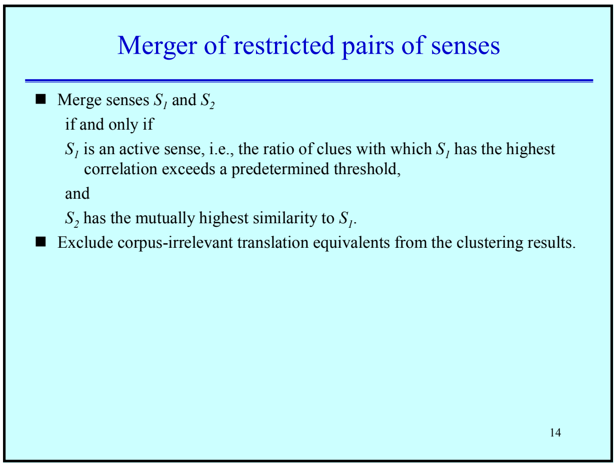
- Distribution pattern subordinate to S_1 :
  Distribution pattern resulting from the sense-vs.-clue correlation matrix for the set of senses excluding sense S_1
- Example

Distribution patterns

Distribution patterns subordinate to {promotion, 宣伝<*SENDEN*>}

- Similarity of sense S_2 to sense S_1:
  Similarity between distribution pattern for S_2 subordinate to S_1 and distribution pattern for S_1

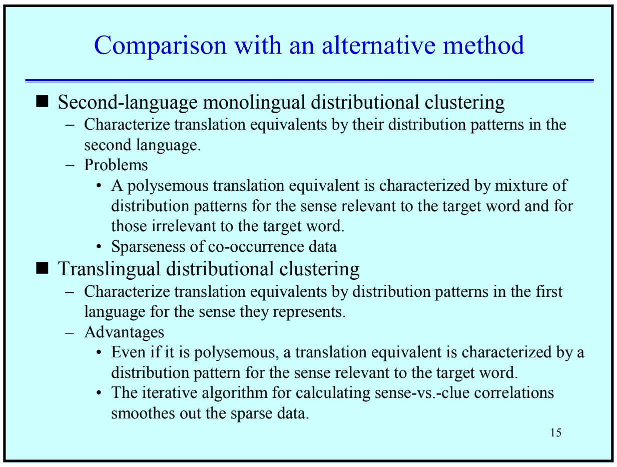
13



Merger of restricted pairs of senses

- Merge senses S_1 and S_2
  if and only if
  S_1 is an active sense, i.e., the ratio of clues with which S_1 has the highest correlation exceeds a predetermined threshold,
  and
  S_2 has the mutually highest similarity to S_1.
- Exclude corpus-irrelevant translation equivalents from the clustering results.

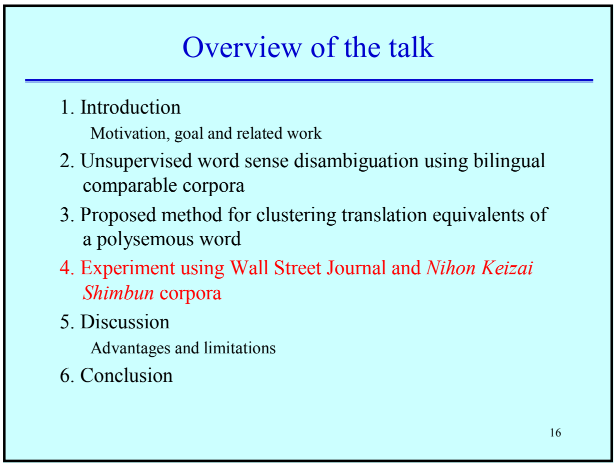
14



Comparison with an alternative method

- Second-language monolingual distributional clustering
  - Characterize translation equivalents by their distribution patterns in the second language.
  - Problems
    - A polysemous translation equivalent is characterized by mixture of distribution patterns for the sense relevant to the target word and for those irrelevant to the target word.
    - Sparseness of co-occurrence data
- Translingual distributional clustering
  - Characterize translation equivalents by distribution patterns in the first language for the sense they represents.
  - Advantages
    - Even if it is polysemous, a translation equivalent is characterized by a distribution pattern for the sense relevant to the target word.
    - The iterative algorithm for calculating sense-vs.-clue correlations smoothes out the sparse data.

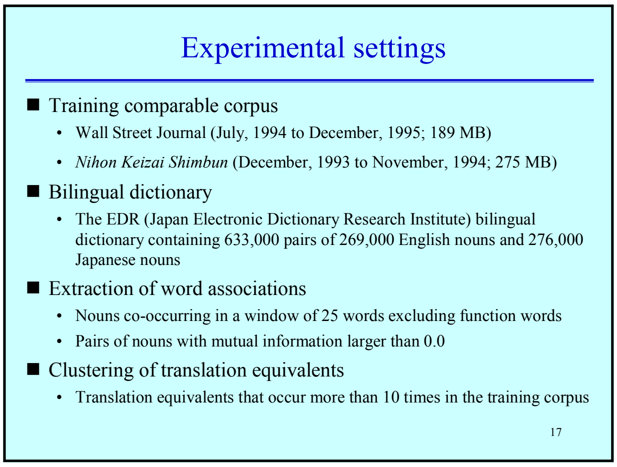
15



Overview of the talk

1. Introduction
   Motivation, goal and related work
2. Unsupervised word sense disambiguation using bilingual comparable corpora
3. Proposed method for clustering translation equivalents of a polysemous word
4. Experiment using Wall Street Journal and *Nihon Keizai Shimbun* corpora
5. Discussion
   Advantages and limitations
6. Conclusion

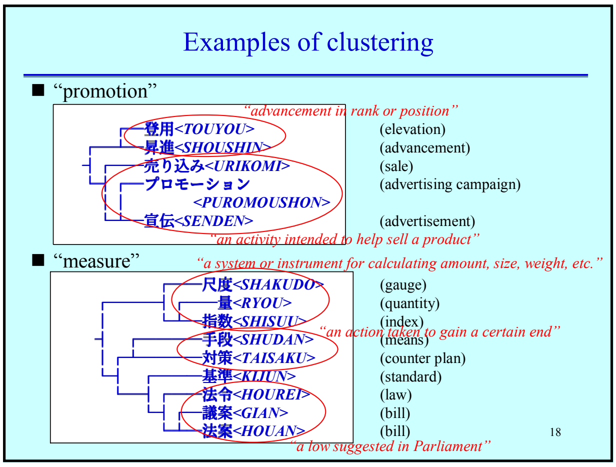
16



Experimental settings

- Training comparable corpus
  - Wall Street Journal (July, 1994 to December, 1995; 189 MB)
  - *Nihon Keizai Shimbun* (December, 1993 to November, 1994; 275 MB)
- Bilingual dictionary
  - The EDR (Japan Electronic Dictionary Research Institute) bilingual dictionary containing 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns
- Extraction of word associations
  - Nouns co-occurring in a window of 25 words excluding function words
  - Pairs of nouns with mutual information larger than 0.0
- Clustering of translation equivalents
  - Translation equivalents that occur more than 10 times in the training corpus

17



Examples of clustering

- “promotion”

“advancement in rank or position”

登用<*TOUYOU*> (elevation)
昇進<*SHOUSHIN*> (advancement)
売り込み<*URIKOMI*> (sale)
プロモーション<*PUROMOUSHON*> (advertising campaign)
宣伝<*SENDEN*> (advertisement)

“an activity intended to help sell a product”

- “measure”

“a system or instrument for calculating amount, size, weight, etc.”

尺度<*SHAKUDO*> (gauge)
量<*RYOU*> (quantity)
指数<*SHISUU*> (index)

“an action taken to gain a certain end”

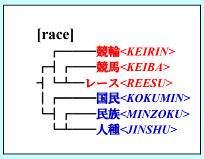
手段<*SHUDAN*> (means)
対策<*TAISAKU*> (counter plan)
基準<*KIJUN*> (standard)
法令<*HOUREI*> (law)
議案<*GIAN*> (bill)
法案<*HOUAN*> (bill)

“a low suggested in Parliament”

18

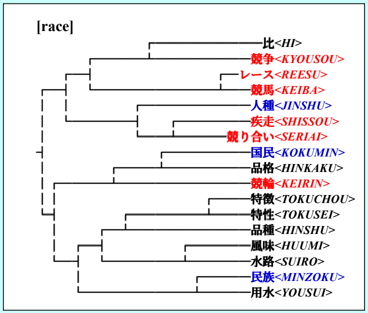
Clustering with proposed method and an alternative method

- Proposed method



[Note]
Red words represent the "competition" sense of "race."
Blue words represent the "group of people" sense of "race."

- Monolingual distributional clustering



Evaluation

- ■ Difficulty in evaluating word sense acquisition methods
  - – Prepare a standard sense inventory
    Translation equivalents used in the standard sense inventory do not always occur frequently in the training corpus.
  - – Establish the complete set of senses appearing in the training corpus.
- ■ Our evaluative measure
  - ● Recall of senses
    Judge a sense as acquired when the output dendrogram includes at least one translation equivalent defining that sense.
  - ● Accuracy of sense definitions
    Regard a set of senses as a set of pairs of translation equivalents defining the same senses, and define the accuracy of sense definitions as the maximum *F*-measure in all cycles of clustering.
- ■ Evaluation results for 60 test English polysemous words
  - ● Recall of senses: 87% for senses whose ratios in the corpus were not less than 5%.
  - ● Accuracy of sense definitions: 77%

5. Discussion

- ■ Advantages of our method
  - ● Corpus-dependent division and definition of word senses
  - ● Unify word sense acquisition with word sense disambiguation—acquire senses of distinguishable granularity.
  - ● Effective for translation equivalents with moderate occurrence frequencies
  - ● Moderate computational load—35 seconds per target polysemous word on a Windows 2000 server (CPU; Pentium 4 (1.9 GHz), memory: 2 GB)
- ■ Limitations and directions for extension
  - ● Difficulty in determining how many senses are appropriate for each target word
    Avoid merging senses having complementary distribution patterns.
  - ● Effective for topical senses but ineffective for generic senses
    Use syntactic co-occurrence.

6. Conclusion

- ■ Unified approach to word sense acquisition and word sense disambiguation using a bilingual comparable corpus and a bilingual dictionary
- ■ An experiment using Wall Street Journal and *Nihon Keizai Shimbun* corpora and the EDR bilingual dictionary demonstrated the effectiveness of the method.