AUTOMATIC DETECTION OF NEAR-SYNONYMS IN NEWS CLUSTERS

Authors: Aleksey Alekseev
Natalia Loukachevitch

NEWS FLOW PROCESSING

- News aggregators (30-40 thousands of documents per day)
- Thematic clustering of news documents – news clusters:
  - Duplicate Removing
  - Thematic Categorization
  - Automatic Summarization
  - Novelty Detection
  - Information Extraction
- Most operations are performed on the basis of word-sequence representation of a text

WORD-SEQUENCE REPRESENTATION PROBLEMS

- Just one entity could be called in text by not only one consequent word (by multiword expression)
- Different names could be used in news clusters for mention of the same real entity
- U.S. air base in Kyrgyzstan:
  - *Manas base, Manas airbase,*
  - *Manas, base at Manas International Airport,*
  - *U.S. base, U.S. air base*
- **Problems:**
  - Cluster boundaries correction
  - Automatic summarization
  - Novelty detection
  - Sub-cluster selection and etc.

METHODS FOR REPRESENTATION COMPLICATION

- Using of Thesauruses
  - Synonyms
  - Phrases
  - Lexical chains
  - **BUT:** it is impossible to describe all the cases in pre-created resource
- Co-reference resolution
  - Named entities
  - Full and partial naming
  - Co-reference resolution
  - BUT: not only named entities have a problem of naming variability

MULTIWORD EXPESSIONS AND NEAR-SYNONYMS

- **Multiword expressions**, parts of which frequently do not reflect the meaning of the whole expression («Russian Federation», «Manas Airbase»)
- **Near-synonyms** – words end phrases, which are not synonyms in general context, but could be interpreted as synonyms in specific context
- For example, "*Parliament*" and "*Parliamentarian*" words are not synonyms, but within a particular cluster, e.g. in which decision-making process in a parliament is discussed, these words may be classified as synonyms, or near-synonyms.
- Multiword expression extraction and near-synonym detection are very important for different areas of Computer Linguistics (Information Retrieval, Summarization and etc.)

RESEARCH BASIS

- A news cluster contains a lot of document on the same thematic
- Rewriters specially reformulate texts by using synonyms and similar meaning words
- **Task:** on the basis of news cluster structure
  - To extract multiword expressions, which denote the main entities of cluster
  - To detect words and phrases, which are synonyms in context of the concrete news cluster
- **Method:**
  - Coherent text properties – global coherence
  - News cluster (is devoted to the just one theme)

GLOBAL TEXT COHERENCE

- **Van Dijk and global coherence hypothesis (1985)**
- A coherent text has one main theme and this theme could be expressed as a proposition
- Whole text theme is revealed in the text by local themes
- Each text sentence corresponds to some text theme
- Global coherence mechanism allows to control local connectives and transitions

LEXICAL COHERENCE vs. GLOBAL COHERENCE

- Coherent text has lexical coherence: lexical and semantic repetitions
- Lexical coherence is an instrument of global coherence
- The more two entities are mentioned in the same sentences, the more their relation is important for whole text content
- If two entities are frequently mentioned in the text, but they rarely occur in the same sentences, it may mean, that they have some meaning relation (e.g. Synonym, Hyponym – Hypernym, Meronym-Holonym)

HYPOTHESIS CHECKING - 1

- Hypothesis checking has been carried out with the help of Russian language Thesaurus RuThes.
- Related in Thesaurus objects have been treated as a positive examples of near-synonymy
- Different types of relations were considered independently
- Two part of speech groups:
  - Noun + Noun
  - Adjective + Noun
- For each pair a number of the same sentence (Fsent) and neighboring sentence (Fsegm) occurrences have been counted

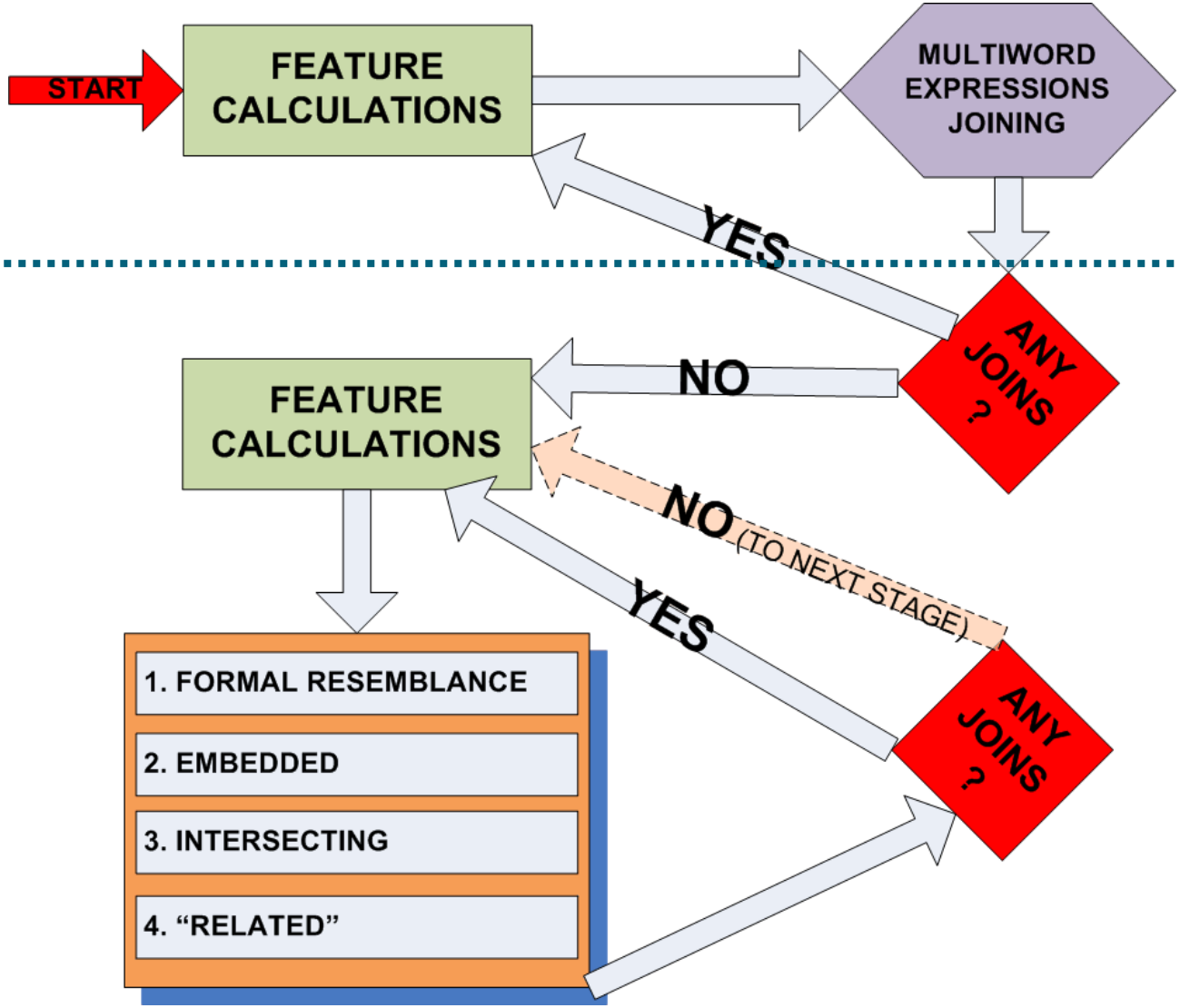
HYPOTHESIS CHECKING - 2

Relation Type	Fsegm / Fsent	Number of pairs
Synonyms (Noun + Noun)	**0.309**	31
Synonyms (Adjective + Noun)	**0.491**	53
Hyponym–Hypernym (Noun + Noun)	1.130	88
Hyponym–Hypernym (Adj. + Noun)	1.471	28
Meronym - Holonym (Noun + Noun)	0.779	58
Meronym - Holonym (Adj. + Noun)	1.580	29
Others	**1.440**	21483

NEWS CLUSTERS AND COHERENT TEXT PROPERTIES

- A news cluster is not a connected text, but:
  - It has the theme of cluster
  - Statistical features are significantly enhanced
- **Goals:**
  - Multiword expression extraction
  - Near-synonym groups detection
- **Example:** news cluster of 19.02.2009
- Theme: Kyrgyzstan and the United States agreement denunciation over U.S. air base located at the Manas International Airport
- 195 news documents

GENERAL METHOD SCHEME



WORD CONTEXTS EXTRACTION

- Sentences are divided into segments between punctuation marks
- Contexts of word W:
  - Neighboring adjectives or nouns situated directly to the right or left from W (*Near*)
  - Adjectives and nouns occurring in sentence segments with a verb, and the verb is located between W and these adjectives or nouns (*Av*),
  - Adjectives and nouns that are not separated with a verb from W and are not direct neighbors to W (*NotN*),
  - Adjectives and nouns which are situated in neighboring sentences with word W (*Ns*).
- Frequencies for all types of contexts are memorized for each pair of words (expressions)

MULTIWORD EXPRESSION EXTRACTION STAGE

- This stage is considered as a required pre-processing stage for near-synonym detection
- Significant excess in co-occurrence frequency of neighbor words in comparison with their separate occurrence frequency in segments of sentences is the main criterion:

  Near > 2 * (Av + NotN)
- In addition, the restrictions on Near feature are imposed:

  Near > 0.4 * MaxAv

  Near > 0.2 * Min (Freq1, Freq2)
- **Examples from the cluster:**
  - *Parliament of Kyrgyzstan,*
  - *the U.S. military, denunciation of agreement with the U.S.,*
  - *Kyrgyz President Kurmanbek Bakiyev*

NEAR-SYNONYM DETECTION

- The following main factors are exploited for assuming a semantic relation between expressions U1 and U2:
  - U1 and U2 co-occur more often in neighboring sentences than within segments of the same sentences,
  - U1 and U2 have similar contexts based on Near, AcrossVerb, NotNear and Ns features, which are determined by calculating scalar products of corresponding vectors (NearScalProd, AVerbScalProd, NotNearScalProd, NsentScalProd)

NEAR-SYNONYM DETECTION STAGES - 1

- There are four independent steps at near-synonym detection stage. Each step is devoted to detection of near-synonyms with special distinguishing characteristics.
- The number of steps could be increased (and it is going to be done in further work), if new sufficiently good features for near-synonym detection would be obtained.
- The steps are executed consequently (in order of precision decreasing), each next step receives the results from the previous step (the first step takes the results of multiword expression extraction stage)

NEAR-SYNONYM DETECTION STAGES - 2

1. **Expressions with formal resemblance**
   - *Kirghizia – Kyrgyz;*
   - *Kyrgyz parliament – Parliament of Kyrgyzstan*
2. **Embedded expressions**
   - *Parliament - Parliament of Kyrgyzstan,*
   - *Air Base – Manas Air Base*
3. **Intercepting expressions with equal length**
   - *Manas Air Base – Manas base,*
   - *American military - U.S. side*
4. **«Arbitrary» expressions**
   - *U.S.A - American*

NEAR-SYNONYM DETECTION ALGORITHM

- Different sets of criteria are applied for each pair type
- The lookup is performed in order of frequency decreasing:
  - for every expression U1, all expressions U2 having a lower frequency than U1, are considered
- If all conditions are satisfied, then:
  - less frequent expression U2 is postulated as a synonym of U1 expression
  - all U2 contexts are transferred to U1 contexts, the expressions U1 and U2 become joined together
- The process is iterative (until at least one join was performed)
- As a result, the sets of near-synonyms (synonym groups) with selected main expression are produced

CONDITION EXAMPLES ON EXPRESSION COMPARISON

- High co-occurrence frequency in neighboring sentences
  - *NS > 2 * (Av + Near + NotN)*
  - *NS > 0.1 * MaxAv*
- Occurrence frequency constraints:
  - *Freq1, Freq2 > 0.5 * MaxAv*
- Similarity of neighboring contexts:
  - *NearScalProd > 0.3*
- Join instances from the example cluster
  - *U.S.A – American*
  - *Kirghizia - Bishkek*

EXTRACTED SYNONYM GROUPS FOR EXAMPLE CLUSTER

- ***Manas base***: *base, Manas Air Base, Air Base, Manas;*
- ***USA***: *American, America;*
- ***Kyrgyzstan***: *Kirghizia, Kyrgyz, Kyrgyz-American, Bishkek;*
- ***Parliament of Kyrgyzstan***: *Kyrgyz parliament, parliament, parliamentary, parliamentarian;*
- ***Manas International Airport***: *airport, Manas airport;*
- ***Bill***: *law, legislation, legislative, legal* and etc.

METHOD EVALUATION

- 10 news clusters on various topics (sport, politics, incidents)
- More than 40 documents in each cluster

- **Evaluation of quality:**
  - Multiword expression extraction
  - Near-synonym groups formation
    - Semantic relatedness of every synonym in a group to its main synonym evaluation
    - *Parliament of Kyrgyzstan: Kyrgyz Parliament, Parliament, parliamentary, parliamentarian*

EVALUATION OF MULTIWORD EXPRESSION EXTRACTION

- Percentage of syntactically correct groups among all extracted expressions (precision evaluation)
- Professional linguist attraction for the most significant multiword expressions selection (5-10 for each cluster, recall evaluation)
- For the example cluster:
  - *Manas Airbase, Parliament of Kyrgyzstan, Manas base,*
  - *Kyrgyz Parliament, Denunciation of agreement ,*
  - *Government's decision.*
- Automatic extraction results:
  - Correct syntactic groups percentage - 87.9% (from 364)
  - Recall evaluation (70 expressions overall) - 72.6 %

EVALUATION OF NEAR-SYNONYM DETECTION - 1

- Each element of synonym group was evaluated for relatedness to main synonym of the group, during the near-synonym detection evaluation procedure.
- For each element occurrence of each synonym group element was evaluated its correctness in this group in concrete case and the two-value score (“positive” or “negative”) was marked
- If the total number of “positive” marks for estimated element exceeded the total number of “negative” marks, then the join was marked as “positive” in general
- All the steps were evaluated independently

EVALUATION OF NEAR-SYNONYM DETECTION - 2

- 1. Expressions with formal resemblance joining
  - ✓ **87.9% (155)**
- 2. Embedded expressions joining
  - ✓ **91.4% (99)**
- 3. Intercepting expressions with equal length joining
  - ✓ **85.7% (8)**
- 4. «Arbitrary» expressions joining
  - ✓ **62.5% (38)**

CONCLUSION

- Information retrieval method for news cluster has been proposed:
  - Multiword expression extraction
  - Near-synonym groups detection
- The evaluation of performed automatic operations has been carried out
- In future we are going to use extracted near-synonyms for improving overall performance of such operations as:
  - Automatic summarization
  - Cluster boundaries correction
  - Novelty detection
  - Formation of sub-clusters and etc.