Data-driven sense induction for disambiguation and lexical selection in translation

Marianna Apidianaki, University Paris 7

22 October 2008

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?
ii. supervised WSD
iii. automatic sense acquisition
iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus
ii. underlying assumptions and implementation
iii. cross-lingual projection of semantic information
iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method
ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?
ii. supervised WSD
iii. automatic sense acquisition
iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus
ii. underlying assumptions and implementation
iii. cross-lingual projection of semantic information
iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method
ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

i. What is WSD?

What is it?

an **intermediary** stage of processing that aims to ameliorate the performance of NLP applications (Wilks & Stevenson, '96)

What do we need?

- a **sense inventory** describing the senses of ambiguous words
- a **method** that can decide which sense is carried by a new instance

i. What is WSD?

What is it?

an **intermediary** stage of processing that aims to ameliorate the performance of NLP applications (Wilks & Stevenson, '96)

What do we need?

- a **sense inventory** describing the senses of ambiguous words
- a **method** that can decide which sense is carried by a new instance

Supervised methods

- need of a **sense-tagged** corpus (senses taken from a predefined sense inventory)
- learning of **contextual regularities** linked to the senses of the words

Unsupervised methods

- no need of a **sense-tagged** corpus
- exploitation of the results of **automatic sense acquisition** methods

ii. Supervised WSD

Main advantage : the supervised WSD methods perform better than the unsupervised ones

ii. Supervised WSD

Main advantage : the supervised WSD methods perform better than the unsupervised ones

Drawbacks :

- very few sense-tagged corpora
- need of predefined semantic ressources
  - **not available** in many languages
  - **qualitative** and **structural divergences**
  - semantic information not relative to the domains of the processed texts
  - **great number** and **proximity** of **senses**, absence of explicit links (Dolan, '94; Pustejovsky, '95; Edmonds & Kilgarriff, '02)
    - ➢ **WSD algorithms** confronted with multiple *correct* choices → **complex processing** and **selection**
  - **fine granularity** : not needed in some applications (MT, IR) (Mihalcea & Moldovan, '01)
    - ➢ need of adaptation to the WSD requirements of specific applications

ii. Supervised WSD

Main advantage : the supervised WSD methods perform better than the unsupervised ones

Drawbacks :

- very few sense-tagged corpora
- need of predefined semantic ressources
    - **not available** in many languages
    - **qualitative** and **structural divergences**
    - semantic information not relative to the domains of the processed texts
    - **great number** and **proximity** of **senses**, absence of explicit links (Dolan, '94; Pustejovsky, '95; Edmonds & Kilgarriff, '02)
        - **WSD algorithms** confronted with multiple *correct* choices → **complex processing** and **selection**
    - **fine granularity** : not needed in some applications (MT, IR) (Mihalcea & Moldovan, '01)
        - need of adaptation to the WSD requirements of specific applications

=> arguments towards... a. data-driven sense acquisition

b. unsupervised WSD

iii. Data-driven sense acquisition

Monolingual context

- distributional hypothesis of meaning (Harris, '54)
- sense acquisition : an **unsupervised machine learning** problem

iii. Data-driven sense acquisition

Monolingual context

- distributional hypothesis of meaning (Harris, '54)
- sense acquisition : an **unsupervised machine learning** problem

Unsupervised algorithms

- **sense clustering** : grouping of semantically similar instances on the basis of their similar distributional behaviour (Schütze, '92, '98; Pedersen & Bruce, '97; Widdows & Dorow, '02)
- instances of ambiguous words : characterized by the features found in their lexical context (direct or indirect cooccurrences (Pantel & Lin, '02; Véronis, '03; Dorow & Widdows, '03; // Schütze, '98; Ferret, '04))
- construction of a vector or similarity space, or elaboration of cooccurrence graphs
- **distance measure** : determines the way in which the similarity of two elements is calculated. In sense clustering, it corresponds to the similarity of the sets of context features corresponding to different word instances.

Monolingual context

Advantages

- ressource creation for different languages
- senses related to the processed data

Disadvantages

- specificity of the senses to the corpus from which they derive (Pereira *et al*., '93)
- strong impact of the corpus on the coverage of the inventory
- difficult interpretation of the senses
- fine granularity of sense distinctions (uses)
- sensibility to the data sparseness effect (Purandare & Pedersen, '04)

iii. Data-driven sense acquisition

Translation context

Different **lexicalisation** of SL word senses in other languages

→ equivalents (EQVs) : clues for sense distinctions (ex. bank: *banque-rive*, duty: *droit-devoir*)

Translation context

Different **lexicalisation** of SL word senses in other languages

→ equivalents (EQVs) : clues for sense distinctions (ex. bank: *banque-rive*, duty: *droit-devoir*)

Advantages :

- translations : objective source of semantic information (Resnik & Yarowsky, '00)
- automatic creation of sense-tagged corpora
- conformity to bi- (multi-)lingual processing (lexical selection in MT; Ng *et al*. '03)

Eventual problems during SL sense distinction :

- translation ambiguity (Resnik & Yarowski, *ibid*.; Ide *et al*., '02)
- sense distinctions valid only in the TL (Fuchs, '96)
- semantic similarity of the EQVs

iv. Data-driven and application-oriented WSD

Tendency towards unsupervised WSD methods :

- no need for tagged data
- exploited information : results of data-driven sense induction methods

iv. Data-driven and application-oriented WSD

Tendency towards unsupervised WSD methods :

- no need for tagged data
- exploited information : results of data-driven sense induction methods

Tendency towards application-oriented WSD :

- WSD : an intermediary stage of processing (Wilks & Stevenson, '96)
- varying WSD needs in different applications (Resnik & Yarowsky, '97; Mihalcea & Moldovan, '01)
- absence of link between WSD methods and the finality of applications : common criticism

iv. Data-driven and application-oriented WSD

Tendency towards unsupervised WSD methods :

- no need for tagged data
- exploited information : results of data-driven sense induction methods

Tendency towards application-oriented WSD :

- WSD : an intermediary stage of processing (Wilks & Stevenson, '96)
- varying WSD needs in different applications (Resnik & Yarowsky, '97; Mihalcea & Moldovan, '01)
- absence of link between WSD methods and the finality of applications : common criticism

WSD for Translation :

- assimilation of the WSD and lexical selection tasks (Kaji *et al*., '03; Vickrey *et al*., '05; Specia, '05)
- great availability of annotated data in the form of word-aligned parallel corpora
- no need of spotting fine sense distinctions

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?

ii. supervised WSD

iii. automatic sense acquisition

iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus

ii. underlying assumptions and implementation

iii. cross-lingual projection of semantic information

iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method

ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

i. Training corpus

English-Greek part of the **INTERA** parallel corpus (Gavrilidou *et al*., 04)

- POS-tagged, lemmatized, sentence aligned
- 4 000 000 words
- different domains : law (42%), health (24%), education (21%), tourism (11%), environment (2%)

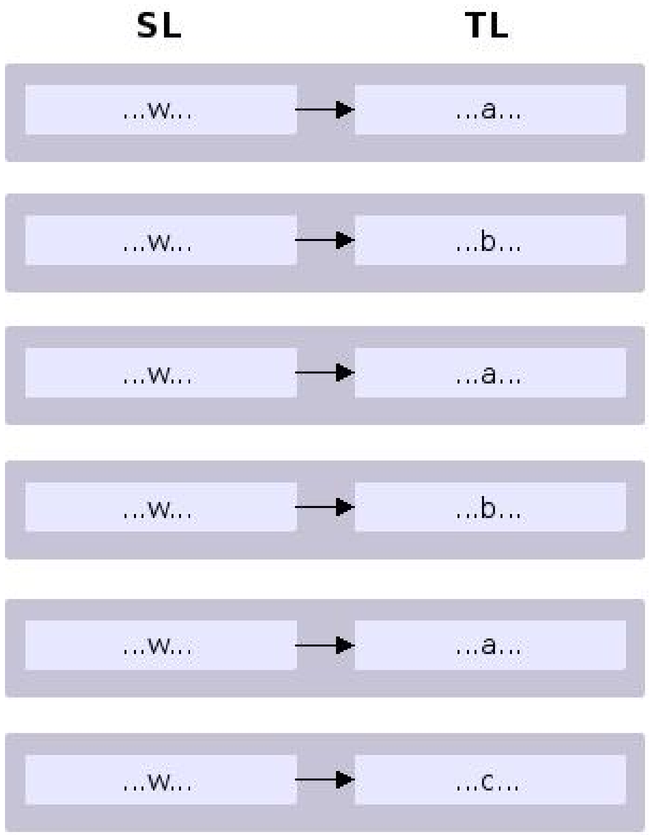
Further preprocessing :

- word alignment (tokens, types)
- bilingual lexicon creation (EN-GR, GR-EN)
- filtering of the lexicons
- manual elaboration of a lexical sample : 150 entries

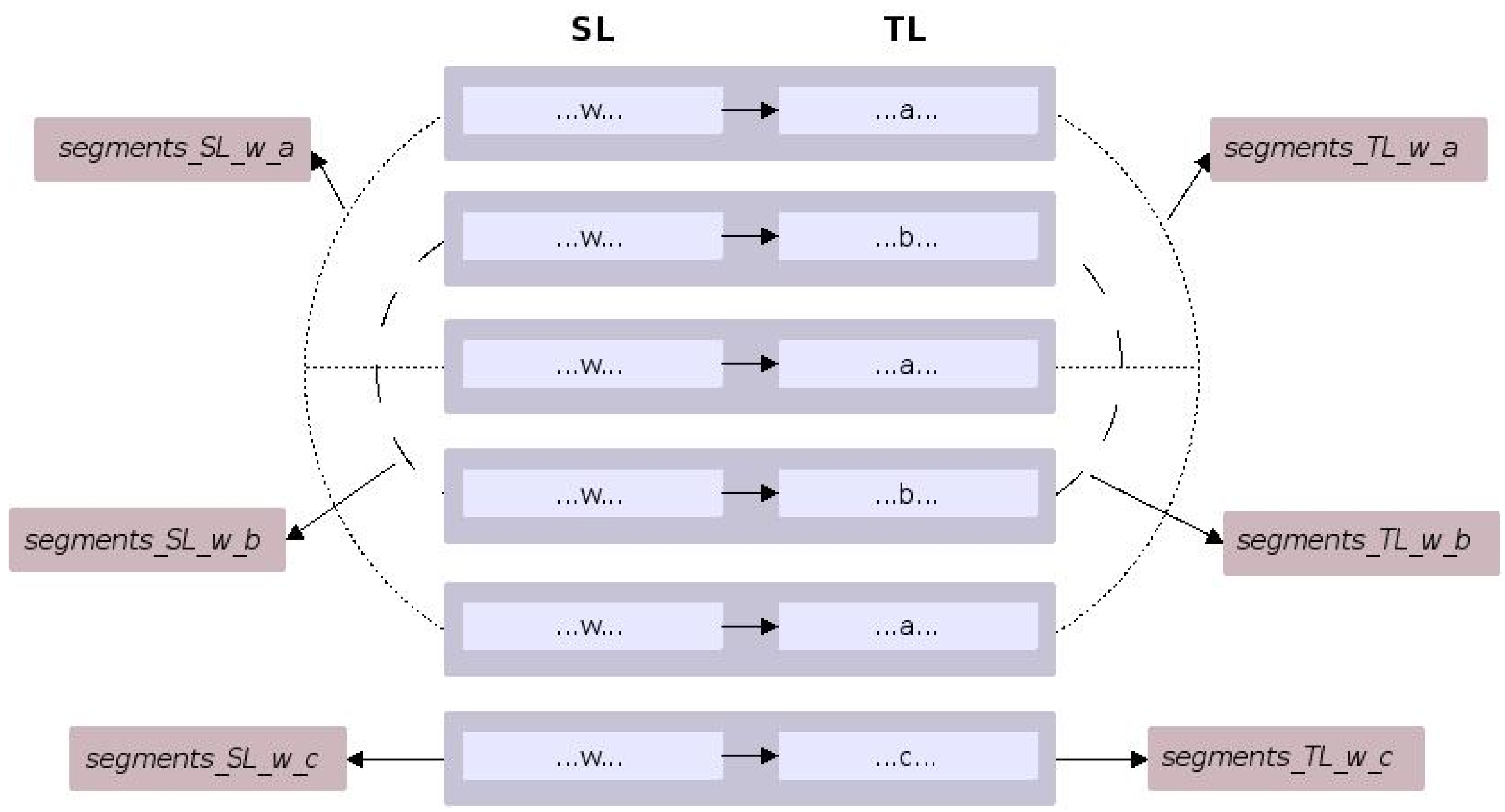
Manual translation spotting (Véronis & Langlais, '00; Simard, '03) : 10 ambiguous words

i. Training corpus

Sub-corpus creation for each ambiguous word (*w*)



Sub-corpora filtering by reference to the translation EQVs



Elaboration of a data-driven sense acquisition method
ii. Underlying assumptions and implementation

Combination of **translation** and **cooccurrence** information coming from a parallel aligned corpus.

Theoretical assumptions

a) distributional hypotheses of meaning (Harris, '54) and of semantic similarity (Charles & Miller, '89)

b) cross-lingual sense correspondance between words in translation relation
(« equivalence in context », Chesterman, '98)

ii. Underlying assumptions and implementation

Combination of **translation** and **cooccurrence** information coming from a parallel aligned corpus.

Theoretical assumptions

a) distributional hypotheses of meaning (Harris, '54) and of semantic similarity (Charles & Miller, '89)

b) cross-lingual sense correspondance between words in translation relation
(« equivalence in context », Chesterman, '98)

c) Information coming from the lexical contexts of the SL word, when translated by a precise EQV, may shed light to the sense(s) translated and, thus, carried by the EQV.

Elaboration of a data-driven sense acquisition method
ii. Underlying assumptions and implementation

Combination of **translation** and **cooccurrence** information coming from a parallel aligned corpus.

Theoretical assumptions

a) distributional hypotheses of meaning (Harris, '54) and of semantic similarity (Charles & Miller, '89)

b) cross-lingual sense correspondance between words in translation relation
(« equivalence in context », Chesterman, '98)

c) Information coming from the lexical contexts of the SL word, when translated by a precise EQV, may shed light to the sense(s) translated and, thus, carried by the EQV.

Unsupervised machine learning

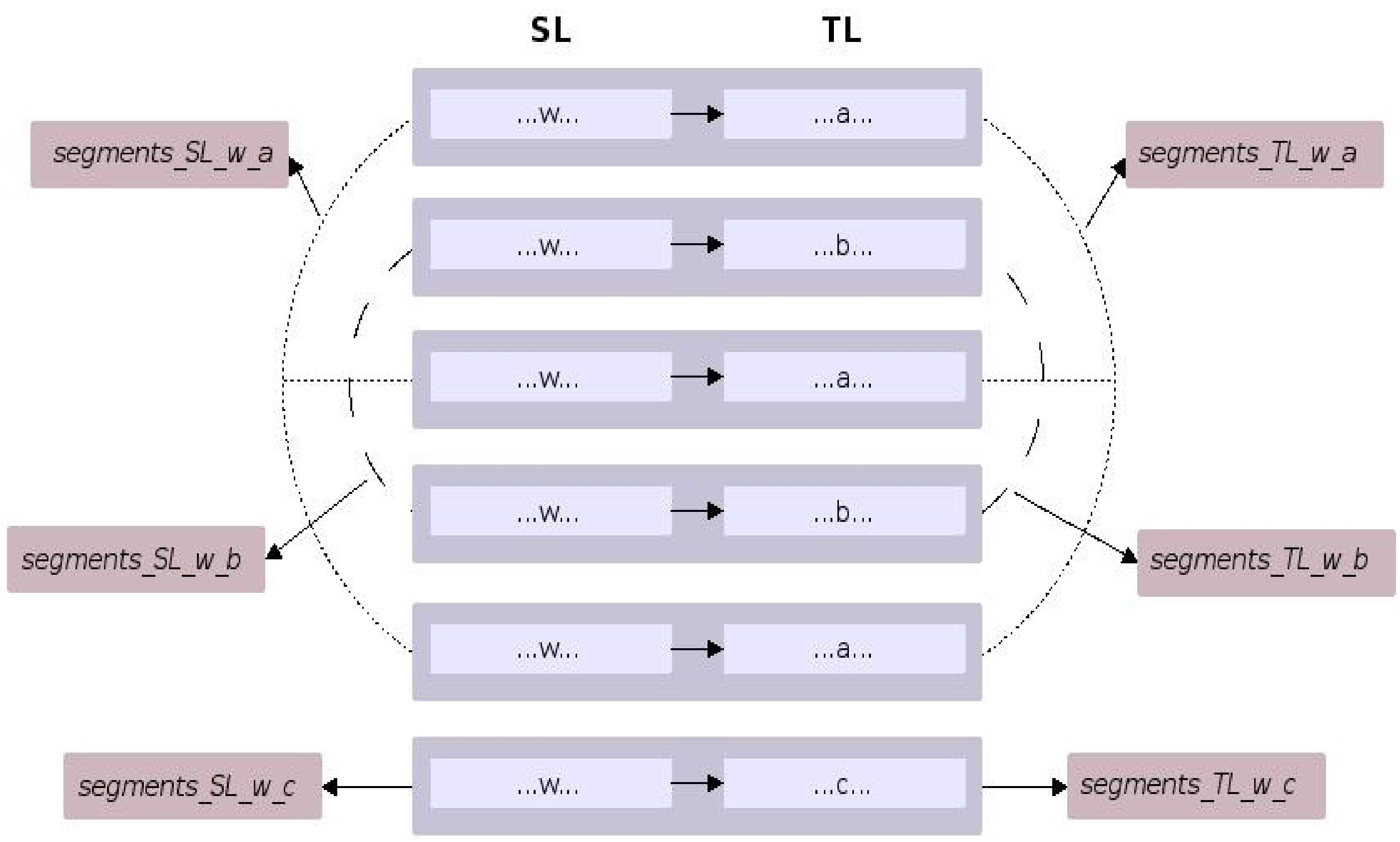
Unsupervised learning algorithms : <u>input</u> → non classified objects
<u>output</u> → groups (clusters) of similar objects

Objects : the EQVs of an ambiguous SL word

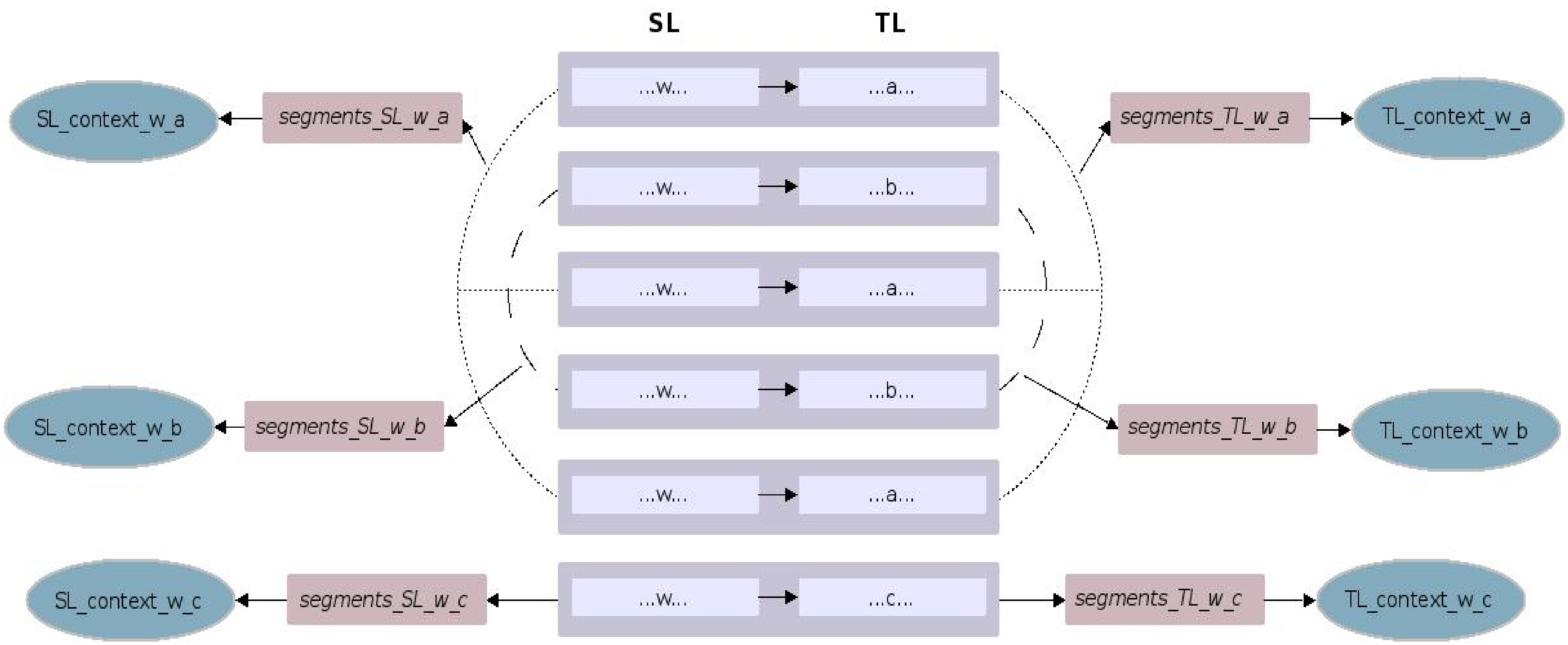
Distance measure : results of a semantic (distributional) similarity calculation in the SL

ii. Underlying assumptions and implementation

Features used for the similarity calculation : the content words of the SL context of each EQV



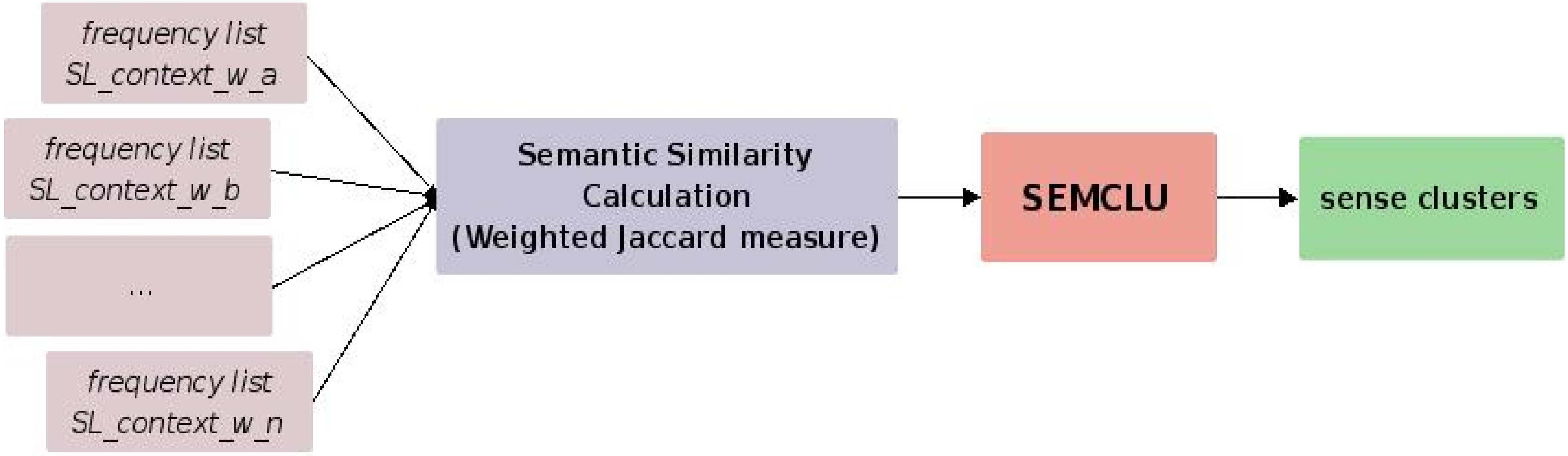
Features used for the similarity calculation : the content words of the SL context of each EQV



Semantic clustering by dynamic programming

Global problem : construction of clusters of semantically similar EQVs (sense clusters)

Sub-problems : estimation of the similarity of pairs of EQVs



movement

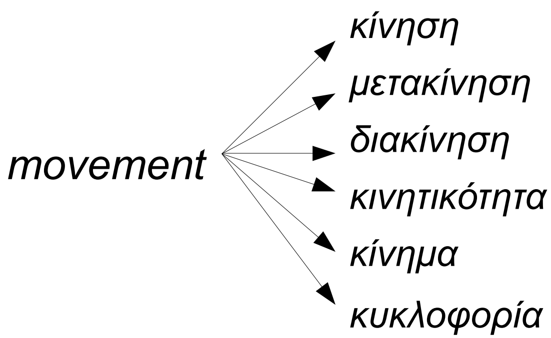
Elaboration of a data-driven sense acquisition method

iii. Cross-lingual projection of semantic information

movement

- *κίνηση*
- *μετακίνηση*
- *διακίνηση*
- *κινητικότητα*
- *κίνημα*
- *κυκλοφορία*

Elaboration of a data-driven sense acquisition method
iii. Cross-lingual projection of semantic information



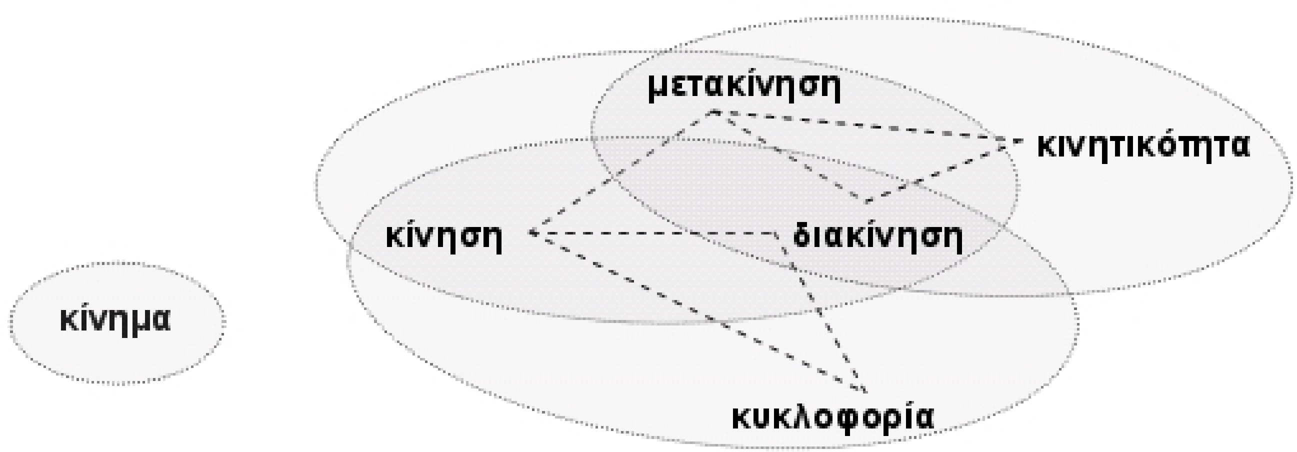
Senses of *movement* :

a. *movement* - {*μετακίνηση*, *κίνηση*, *διακίνηση*}

b. *movement* - {*κίνηση*, *διακίνηση*, *κυκλοφορία*}

c. *movement* - {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}

d. *movement* - {*κίνημα*}



iv. Strengths and weaknesses

Strengths

Processing

- unsupervised method (language-independent)
- data-driven method : senses relevant to the corpus, easy updating of the inventory
- fuzzy clustering

Theoretical level

- distributional hypothesis in a bilingual framework
- differentiation of the senses by reference to their granularity and their proximity
- consideration of parallel ambiguity (EQVs found in the intersection of clusters)
- enrichment of translation correspondances by paradigmatic information

Elaboration of a data-driven sense acquisition method
iv. Strengths and weaknesses

Strengths

- unsupervised method (language-independent)
- data-driven method : senses relevant to the corpus, easy updating of the inventory
- fuzzy clustering

- distributional hypothesis in a bilingual framework
- differentiation of the senses by reference to their granularity and their proximity
- consideration of parallel ambiguity (EQVs found in the intersection of clusters)
- enrichment of translation correspondances by paradigmatic information

Processing

Theoretical level

Weaknesses

- vulnerability to data sparseness (first-order cooccurrences)
- sensibility to the noise present in the alignment results

- analysis of the semantics of the EQVs
- no specification of the relations between clustered EQVs
- risks inherent in the construction of coarse-grained senses

Processing

Theoretical level

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?

ii. supervised WSD

iii. automatic sense acquisition

iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus

ii. underlying assumptions and implementation

iii. cross-lingual projection of semantic information

iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method

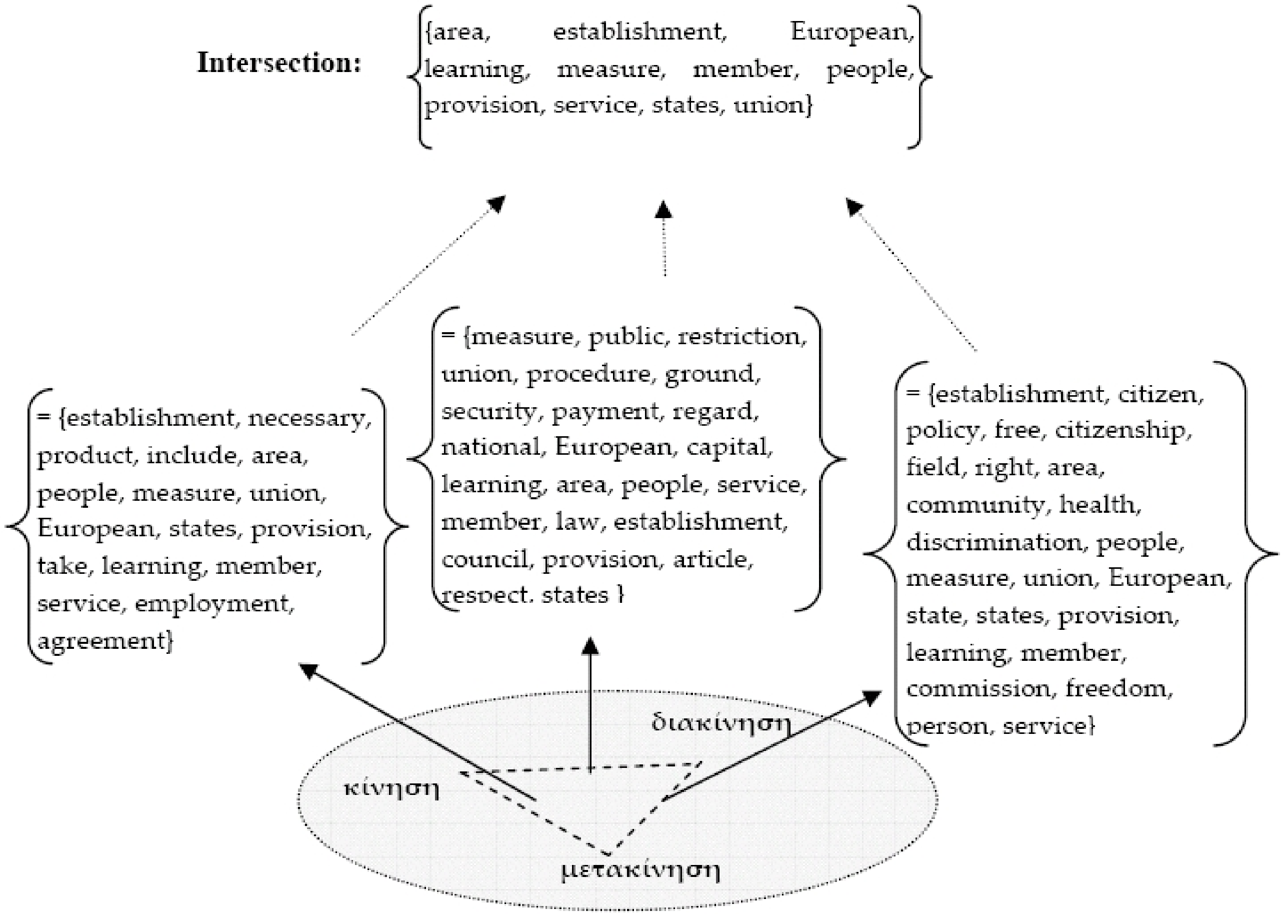
ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

WSD based on the semantic clustering

Information acquired during training

The contextual information that revealed the clustered EQVs' similarity relations characterize the generated clusters.



WSD based on the semantic clustering

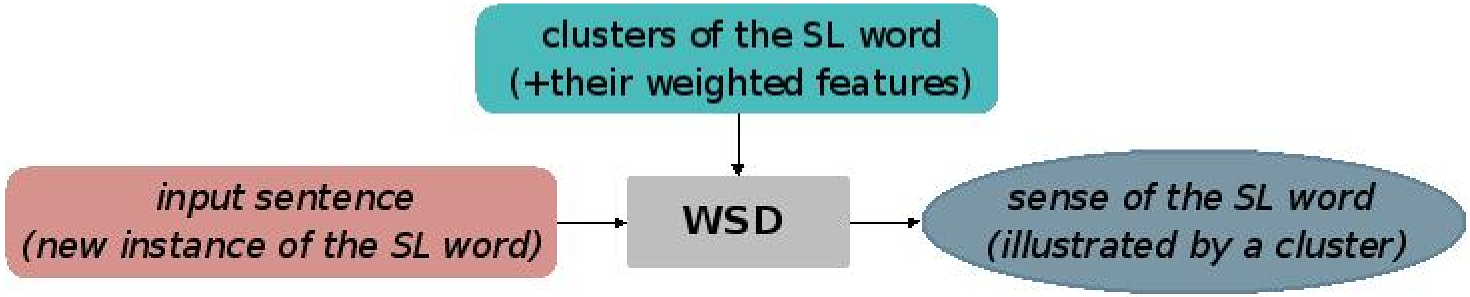
Contextual Information used for WSD

The cooccurrences of the ambiguous word in the input sentence (lemmatised and POS-tagged)

On the internal market there has been a standstill on many issues, from the free ***movement*** *of persons to the European company statute, to taxation, to the banking and insurance sector.*

{*internal* (JJ), *market* (NN), *have* (V), *be* (V), *standstill* (NN), *many* (JJ), *issue* (NN), *free* (JJ), *person* (NN), *European* (JJ), *company* (NN), *statute* (NN), *taxation* (NN), *banking* (NN), *insurance* (NN), *sector* (NN)}

- comparison of the contextual information to the information characterizing each cluster
- calculation of the **weighted intersection** of the two sets of context features



Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?

ii. supervised WSD

iii. automatic sense acquisition

iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus

ii. underlying assumptions and implementation

iii. cross-lingual projection of semantic information

iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method

ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

WSD-dependent lexical selection in Translation

Intervenes only when the WSD prediction concerns a cluster of **more than one EQVs** :

- more or less substitutable translations of the SL word but maybe not substitutable in the translation.

WSD-dependent lexical selection in Translation

Intervenes only when the WSD prediction concerns a cluster of **more than one EQVs** :

- → more or less substitutable translations of the SL word but maybe not substitutable in the translation.

Information acquired during training

Differentiating TL contexts : acquired during the calculation of the similarity of the EQVs on the basis of their TL contexts

WSD-dependent lexical selection in Translation

Intervenes only when the WSD prediction concerns a cluster of **more than one EQVs** :

- more or less substitutable translations of the SL word but maybe not substitutable in the translation.

Information acquired during training

Differentiating TL contexts : acquired during the calculation of the similarity of the EQVs on the basis of their TL contexts

Contextual information used for lexical selection

- test corpus : the EN-GR part of EUROPARL (Koehn, '05)
- test subcorpus of an ambiguous word : translation units sorted by reference to the EQVs
- **reference translation** : replaced by a blank
- **translation context** : cooccurrences of the blank in the TL sentence

Goal of lexical selection :

resolve a simplified translation problem (Vickrey *et al*., 2005) : **blank-filling**

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?
ii. supervised WSD
iii. automatic sense acquisition
iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus
ii. underlying assumptions and implementation
iii. cross-lingual projection of semantic information
iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method
ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

i. Qualitative evaluation of the acquired senses

Semantic Mirrors

A translation-based semantic analysis method (Dyvik, '98, '03, '05) :

- application to our training data
- creation of a semantic thesaurus

Results : - similarity of the acquired sense descriptions
- consolidation of relations between clustered EQVs and of the grouping of the clusters
- analysis of the ambiguity of the EQVs

BalkaNet

A multilingual ressource where concepts are organized in semantic taxonomies and linked via an *Interlingual Index* (ILI)

Advantages of our method :

- data-driven
- consideration of the status and relations between senses
- possibility of automatic modification of the granularity of senses (BalkaNet : too fine-grained)

ii. Quantitative evaluation of the WSD method

Senseval multilingual tasks (Ckhlovski *et al*., '04) : the translation of an ambiguous word in the test corpus is its **sense tag**

here : reference translation : sense tag of the SL word (points to a sense described by a cluster)
goal : predict the sense carried by the sense tag

Evaluation principles : the proposed sense is **correct** (false) if

- cluster of 1 EQV and the EQV corresponds (does not correspond) to the reference
- cluster of >1 EQVs and the reference is (not) found in the cluster

ii. Quantitative evaluation of the WSD method

Recall = correct predictions / new instances

Precision = correct predictions / predictions made by the system

f-measure = 2 * (precision * recall) / precision + recall

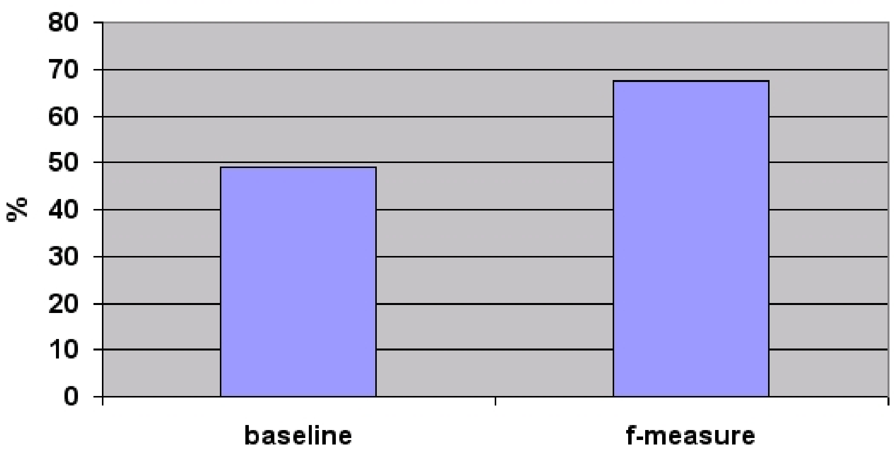
Baseline method

Senseval : the most frequent sense of an ambiguous word (powerful heuristic : asymmetry)

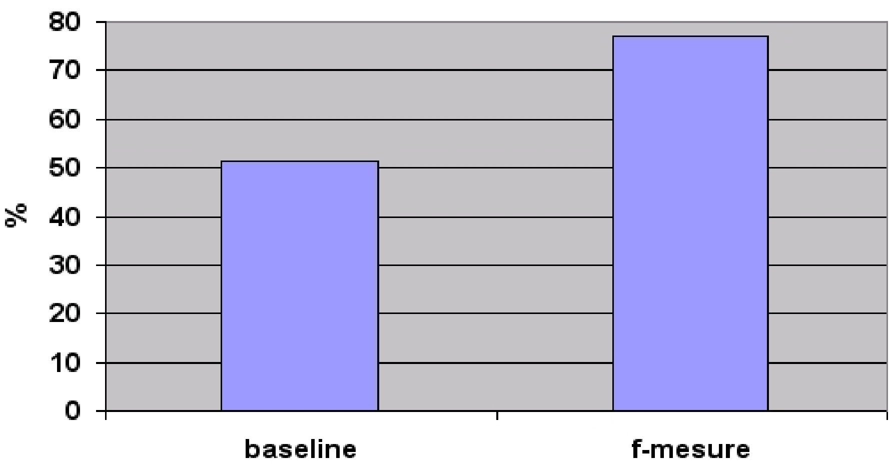
Our baseline : the most frequent EQV in the training corpus (asymmetry)

Baseline score : recall & precision (number of predictions = number of new instances)

Manually created lexicon



Automatically created lexicon



→ the use of the clusters significantly ameliorates the performance of the WSD method

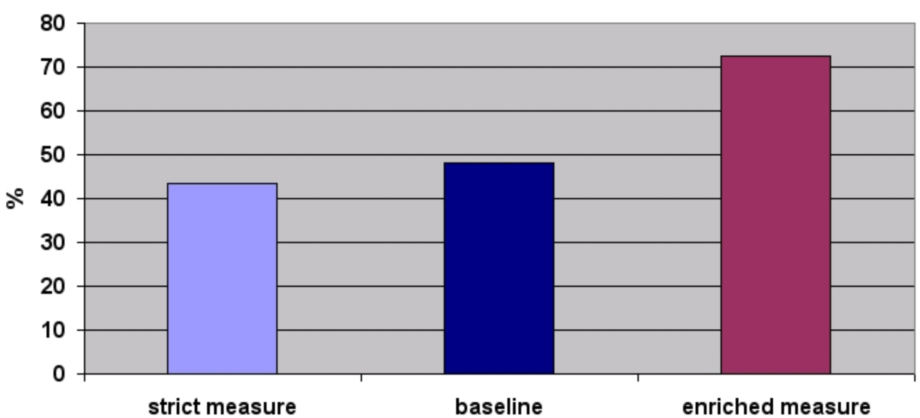
Evaluation
iii. Quantitative evaluation of the lexical selection method

strict precision : only the predictions corresponding exactly to the reference are correct

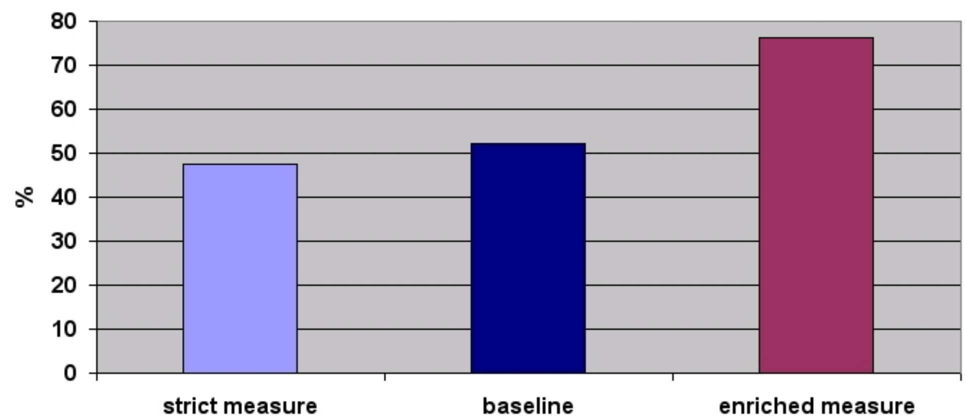
enriched precision : the predictions semantically similar to the reference (found in the same cluster) are correct too

baseline : the most frequent EQV of an ambiguous word

Manually created lexicon :



Automatically created lexicon :



- flexible evaluation (≠ other MT evaluation metrics (Cabezas et Resnik, '05; Callison-Burch *et al.*, '06))
- no need of predefined ressources (METEOR, Banerjee & Lavie, '05; Lavie & Agarwal, '07)
- language-independency : semantic relations automatically identified

Plan of the presentation

a. Towards data-driven sense acquisition and Word Sense Disambiguation (WSD)

i. what is WSD?

ii. supervised WSD

iii. automatic sense acquisition

iv. data-driven and application-oriented WSD

b. Elaboration of a data-driven sense acquisition method

i. training corpus

ii. underlying assumptions and implementation

iii. cross-lingual projection of semantic information

iv. strengths and weaknesses

c. Word Sense Disambiguation based on the semantic clustering

d. WSD-dependent lexical selection in Translation

e. Evaluation

i. **qualititative evaluation** of the sense acquisition method

ii. **quantitative evaluation** of the WSD and the lexical selection methods

f. Conclusion

Conclusion

1. by extending the distributional hypothesis in a bilingual context (and considering translation information) we can automatically induce source language word senses

2. the sense induction process : language-independent

3. construction of sense inventories for languages where such ressources are not available

4. the results of this sense induction process are of benefit for WSD and lexical selection in translation applications
    - amelioration of the performance of the WSD
    - considerable increase of the quantity of semantically pertinent translation predictions

Conclusion

1. by extending the distributional hypothesis in a bilingual context (and considering translation information) we can automatically induce source language word senses
2. the sense induction process : language-independent
3. construction of sense inventories for languages where such ressources are not available
4. the results of this sense induction process are of benefit for WSD and lexical selection in translation applications
   - amelioration of the performance of the WSD
   - considerable increase of the quantity of semantically pertinent translation predictions

Perspectives

1. integration of the WSD method in a SMT system
2. elaboration of an evaluation metric for MT based on the notion of *enriched precision* and application to a more complete task
3. automatic creation of sense-tagged corpora
4. application to other pairs of languages

Thank you