An algorithm for cross-lingual sense clustering tested in a MT evaluation setting

Marianna Apidianaki **Yifan He**

Alpage, INRIA Paris, France

CNGL, DCU Dublin, Ireland

7th International Workshop on Spoken Language Translation (IWSLT)
3 December 2010, Paris

Outline

Outline

Motivation

Semantic resources are needed for Word Sense Disambiguation, semantic similarity calculation, paraphrasing, and other NLP tasks.

Predefined resources

- limited availability and coverage
- too fine granularity for automatic processing (Ide and Wilks, 2007)

Unsupervised sense induction methods

- automatic creation of semantic resources relevant to the domains of interest
- language-independent
- application-oriented

Unsupervised sense induction

Monolingual setting

clustering based on distributional similarities without need for classified training data (Pedersen & Bruce,'97; Schütze,'98, Pantel & Lin,'02)

Multilingual setting

- translations : sense indicators (Cabezas & Resnik,'05;Carpuat & Wu,'07)
- semantic distinctions adapted to the needs of multilingual applications (Resnik & Yarowsky,'00)

Unsupervised sense induction

Monolingual setting

clustering based on distributional similarities without need for classified training data (Pedersen & Bruce,'97; Schütze,'98, Pantel & Lin,'02)

Multilingual setting

- translations : sense indicators (Cabezas & Resnik,'05;Carpuat & Wu,'07)
- semantic distinctions adapted to the needs of multilingual applications (Resnik & Yarowsky,'00)

- no description of semantic relations / parallel ambiguities
- induction of uniform senses => problems in multilingual WSD and its evaluation (Apidianaki,'09)

Unsupervised sense induction

- We propose a **cross-lingual sense induction algorithm** that builds large-scale bilingual semantic resources from parallel corpora where
  - **TEs of a word** (w) : not considered as straightforward indicators of its senses but **clustered** according to their **semantic similarity**
  - **senses of *w*** : described by **sense-clusters** of its semantically similar TEs

Unsupervised sense induction

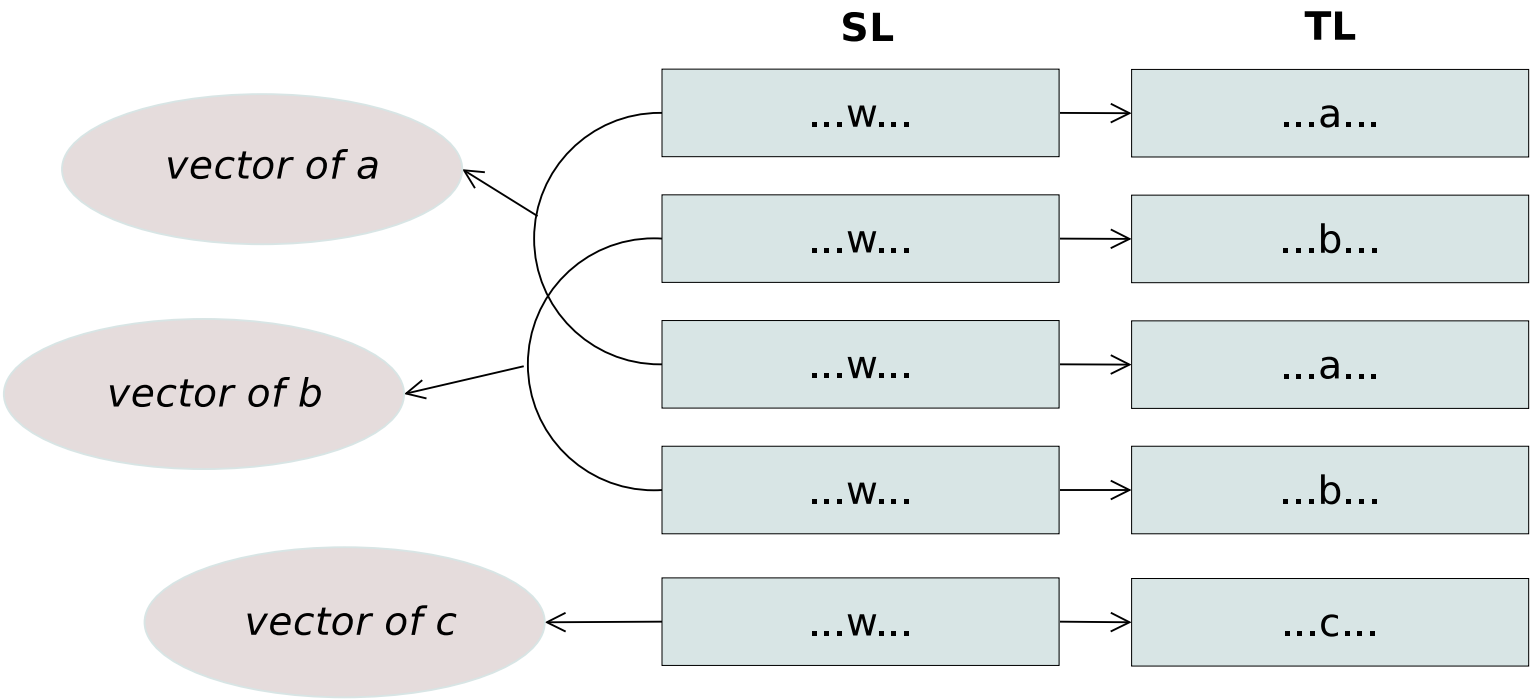
- We propose a **cross-lingual sense induction algorithm** that builds large-scale bilingual semantic resources from parallel corpora where
  - **TEs of a word** (w) : not considered as straightforward indicators of its senses but **clustered** according to their **semantic similarity**
  - **senses of *w*** : described by **sense-clusters** of its semantically similar TEs
- We evaluate the clustering algorithm by integrating the **sense-cluster inventories** in the **METEOR MT evaluation** metric (Lavie and Agarwal,'07).

Outline

Training corpus

EN–FR part of Europal (Koehn,'05) (release v5 containing 1,723,705 sentence pairs)

- elimination of sentence pairs with a great difference in length
- lemmatization, POS-tagging (Schmid,'94)
- word alignment with Giza++ (Och and Ney,'03)
- bilingual lexicons (EN–FR / FR–EN)
- filters : probability score (0.001), intersection, POS (N,V,Adjs)
- for each SL word *w*
  - we keep the TEs translating it >10 times in the training corpus
  - we find the sentences where *w* occurs and where it is translated by each of its TEs
  - we construct a **feature vector** for **each TE** : content words that cooccur with *w* in the sentences where it is translated by the TE


SL
TL
...w...
...a...
...w...
...b...
...w...
...a...
...w...
...b...
...w...
...c...
vector of a
vector of b
vector of c

- **pairwise similarity calculation** of the TEs by a variation of the **Weighted Jaccard** measure (Grefenstette,'94; Apidianaki,'08)
- assignment of a similarity score to each 'TE_pair'

ex. **entretien** : *meeting, talk, maintenance, interview, discussion, conversation*

TE_pairs	score	TE_pairs	score
discussion–meeting	0.328	conversation–talk	0.207
discussion–talk	0.324	interview–talk	0.202
meeting–talk	0.307	maintenance–talk	0.156
conversation–meeting	0.260	maintenance–discussion	0.142
conversation–discussion	0.233	maintenance–interview	0.113
interview–discussion	0.222	maintenance–meeting	0.108
interview–meeting	0.217	conversation–maintenance	0.073
conversation–interview	0.214		

Dynamic thresholding

Threshold defined locally for each *w* and showing the pertinence of the relation of each 'TE_pair'.

1. initial threshold (T) : the mean of the scores (above 0) of the TE_pairs of *w*
2. segmentation of the set of TE_pairs into pairs where score>threshold and pairs where score<threshold (G1, G2)
3. computation of the average of each set (*m1* = average value of G1, *m2* = average value of G2)
4. creation of a new threshold which is the average of *m1* and *m2* (*T* = (m1 + m2)/2)
5. back to step 2, now using the new threshold computed in step 4. Keep repeating until convergence has been reached

ex. **entretien** : *meeting, talk, maintenance, interview, discussion, conversation*

TE_pairs	score	TE_pairs	score
discussion–meeting	**0.328**	conversation–talk	**0.207**
discussion–talk	**0.324**	interview–talk	**0.202**
meeting–talk	**0.307**	maintenance–talk	0.156
conversation–meeting	**0.260**	maintenance–discussion	0.142
conversation–discussion	**0.233**	maintenance–interview	0.113
interview–discussion	**0.222**	maintenance–meeting	0.108
interview–meeting	**0.217**	conversation–maintenance	0.073
conversation–interview	**0.214**		

Threshold : 0.184

The SEMCLU algorithm

Input

- the list of *w*'s TEs (TE_list)
- their **similarity table**
- the **similarity threshold**

Clustering steps

1. each TE_pair with a similarity score over the threshold => **initial cluster** (C)
2. enrichment of the 2-element clusters by additional TEs

- **enrichment condition** : a TE is included in a cluster if it has **pertinent relations** with **all the TEs** already in the cluster (enrichment performed by a recursive function)
- **termination condition** : all the TEs of a *w* are included in some cluster and all their relations have been checked
- **final clusters** : characterized by **global connectivity**
- **1-element clusters** : TEs with no pertinent relation to any other TE of *w*

ex. **entretien** : *meeting, talk, maintenance, interview, discussion, conversation*

TE_pairs	score	TE_pairs	score
discussion–meeting	**0.328**	conversation–talk	**0.207**
discussion–talk	**0.324**	interview–talk	**0.202**
meeting–talk	**0.307**	maintenance–talk	0.156
conversation–meeting	**0.260**	maintenance–discussion	0.142
conversation–discussion	**0.233**	maintenance–interview	0.113
interview–discussion	**0.222**	maintenance–meeting	0.108
interview–meeting	**0.217**	conversation–maintenance	0.073
conversation–interview	**0.214**		

- senses of **entretien**
  1. discussion, meeting, conversation, interview, talk
  2. maintenance

Contents of the sense-cluster inventories

EN–FR

- 3,737 EN words (Nouns : 2,166, Adjectives : 581, Verbs : 990)
- 13,388 clusters (2.66 elements in average)

FR–EN

- 3,734 FR words (Nouns : 2,190, Adjectives : 572, Verbs : 972)
- 11,775 clusters (4.07 elements in average)

Entries from the sense-cluster inventories

Language	POS	*w*	Sense-clusters
EN–FR	**Nouns**	maintenance	{entretien, maintenance} {entretien, maintien} {préservation}
	Verbs	accommodate	{adapter, répondre} {accueillir} {satisfaire, répondre}
	Adjs	tough	{dur, âpre} {dur, sévère, ferme, strict} {rude} {fort}
FR–EN	**Nouns**	limitation	{restriction, limitation, limit} {reduction}
	Verbs	aider	{assist, aid, support} {enable} {aid, help, assist} {contribute}
	Adjs	épineux	{tricky, difficult} {tricky, sensitive} {thorny, difficult}

Outline

Intrinsic evaluation

- lack of a gold standard in lexical semantics
- inadequacy of existing semantic resources for WSD in NLP applications (Edmonds & Kilgarriff,'02; Ide & Wilks,'07)
- varying WSD needs of the applications
- results not meaningful for the usefulness of the inventory in different settings
- unsupervised semantic analysis methods : often application-oriented

Intrinsic evaluation

- lack of a gold standard in lexical semantics
- inadequacy of existing semantic resources for WSD in NLP applications (Edmonds & Kilgarriff,'02; Ide & Wilks,'07)
- varying WSD needs of the applications
- results not meaningful for the usefulness of the inventory in different settings
- unsupervised semantic analysis methods : often application-oriented

Extrinsic evaluation

- integration of the sense-cluster inventories into METEOR
- estimation of the impact of their use during MT evaluation

Semantics-sensitive MT evaluation

- **precision-based metrics** => exact surface correspondences between the compared translations
- attempts to increase the **correlation** of the metrics with **human judgments** of translation quality
- identification of **lexical variation** between hypotheses and reference
  - multiple reference translations (Papineni *et al.*,'02)
  - syntactic structure & dependency information to identify deeper correspondences between sentences (Owczarzak *et al.*,'07)
  - paraphrase detection (Zhou *et al.*,'06; Snover *et al.*,'09)
  - textual entailment (Pado *et al.*,'09)
  - flexible matching of unigrams (METEOR (Lavie & Agarwal,'07))

Evaluation in French

- Exploitation of the FR **sense-cluster** (SC) inventory for rendering METEOR's synonymy module operable for MT evaluation in French.

Evaluation in French

- Exploitation of the FR **sense-cluster** (SC) inventory for rendering METEOR's synonymy module operable for MT evaluation in French.
- Evaluation of the submissions to the **WMT09 English–French shared task**, with and without exploiting the SC inventory (i.e. performing only *exact* and *porter_stemmer* match).

	METEOR	METEOR_SC
Matches	30477.92	32320.25
Chunks	17123.17	18157.83
Score	0.1250	0.1326

Sense correspondences during FR evaluation

Hypothesis	Reference
Tout comme le même prix de l'essence à la pompe des stations est sûrement le produit d'une coïncidence et de la difficile lutte concurrentielle.	De la même manière, un prix de l'essence strictement identique à toutes les stations de service est surement dû au hasard et au dur combat de la concurrence.
Si le voyageur doit confirmer la validité des coupons enregistré, ...	Au cas oú le passager aurait besoin de vérifier la validité des coupons enregistrés, ...
Ils devront abandonner leur relation actuelle avec l'Iran; ...	Ils vont devoir renoncer à leurs relations actuelles avec l'Iran; ...
... et emprisonnés dans la prison locale à 7.00pm, d'être libérés quelques heures plus tard.	... et a été incarcérée vers 19.00 heures et relâchée quelques heures plus tard.

Correlation with human judgments

- **WMT09** evaluation shared task dataset (translations of news stories)
- all **EN–FR human rankings** (390 in total) distributed during this shared evaluation task are used
- correlation of the metrics with human judgments of translation quality measured using the **Spearman's rank order correlation coefficient** (Callison-Burch *et al.*,'08)

$$\rho = 1 - \left(\frac{6 \sum d^2}{n(n^2 - 1)}\right) \quad (1)$$

Segment level correlation of METEOR

	Correlation
BLEU	0.3010±0.0481
METEOR	0.3477±0.0575
METEOR_SC	0.3562±0.0562

Segment level correlation of METEOR

	Correlation
BLEU	0.3010±0.0481
METEOR	0.3477±0.0575
METEOR_SC	0.3562±0.0562

System level correlation of METEOR

	Correlation
BLEU	0.8462
METEOR	0.9021
METEOR_SC	0.9161

Evaluation in English

- Replacement of **WordNet** in METEOR by the **EN sense-cluster** inventory and comparison of the results.

Evaluation in English

- Replacement of **WordNet** in METEOR by the **EN sense-cluster** inventory and comparison of the results.
- **Metrics MATR 2008** development set (human-assigned *adequacy* scores to 1992 sentences generated by 8 MT systems)
- **Source data** : 25 Arabic (AR) language newswire documents with a total of 249 segments (data in each segment: 4 human reference translations in EN and translations from 8 different MT systems)
- **exact**, **stemmed**, **synonymy** and **paraphrase matching**
- **3 types of synonymy** : WordNet (**WN**), SenseClusters (**SC**) and their combination (**WN+SC**)

Evaluation in English

	WN	SC	WN+SC
Synonyms	267.13	356.25	498.78
Paraphrases	370.00	295.88	235.13
Score	0.7123	0.7140	0.7225

Evaluation in English

	WN	SC	WN+SC
Synonyms	267.13	356.25	498.78
Paraphrases	370.00	295.88	235.13
Score	0.7123	0.7140	0.7225

- paraphrases -> larger improvement on this data set
- **SC** : derived from EN–FR bilingual data / **METEOR's paraphrases** : extracted from **TERp paraphrases** induced from AR–EN data (Callison-Burch *et al.*,'08; Snover *et al.*,'09)
- bilingually-motivated linguistic resources work best on the language pairs they are trained on

Sense correspondences during EN evaluation

Hypothesis	Reference
According to Bush, the project should address the main causes of financial crisis and help stabilize the entire economy.	According to Bush, the plan would tackle the basic causes of the financial crisis and help stabilize the entire economy.
The plan supports the financial system will be examined Monday in the House of Representatives.	The plan to support the financial system will be discussed in the House of Representatives on Monday.
"We worked very hard on the issue and have made significant progress toward an agreement that will work and be effective for the market and to all Americans," said Paulson.	"We've worked very hard on this and we've made great progress toward an agreement that will work and that will be useful for all Americans," Paulson said.

Correlation results on the Metrics MATR 2008

Pearson's correlation with human judgments on adequacy.

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{X}}{s_X}\right)\left(\frac{y_i - \bar{Y}}{s_Y}\right) \quad (2)$$

AR–EN segment level correlation

	Correlation	95% Interval
METEOR	0.7256	(0.704, 0.746)
WN-P	0.7455	(0.725, 0.764)
SC-P	0.7442	(0.724, 0.763)
WN+SC-P	0.7526	(0.733, 0.771)
WN+P	0.7634	(0.745, 0.781)
SC+P	0.7614	(0.742, 0.779)
WN+SC+P	0.7657	(0.747, 0.783)

Correlation results on the Metrics MATR 2008

AR–EN system level correlation

	Correlation	95% Interval
METEOR	0.96114	(0.795, 0.993)
WN-P	0.96647	(0.821, 0.994)
SC-P	0.96683	(0.823, 0.994)
WN+SC-P	0.96940	(0.835, 0.995)
WN+P	0.97630	(0.871, 0.996)
SC+P	0.97674	(0.873, 0.996)
WN+SC+P	0.97787	(0.879, 0.996)

Outline

Conclusions

- Positive impact of using the sense-clusters in MT evaluation
- Good quality of the SC resources

Conclusions

- Positive impact of using the sense-clusters in MT evaluation
- Good quality of the SC resources

Evaluation in French

- **synonymy module** of **METEOR** used for evaluation
- identification of matches that would otherwise by missed
- **increase** in the **correlation** of the metric with human judgments

Conclusions

Evaluation in English

- **SC** capture **lexical variation** to an extent comparable to the one observed when a large hand-crafted resource (WordNet) is used
- different size and coverage :
  EN SC base : 11,775 SC / **WordNet** : 664,679 synsets (redundancy, information irrelevant to the domains of interest)
- **training** and **evaluation** on corpora of different domains (Europarl/news stories)

Future work

- generation of SC inventories from **different data sets**
- **language independency** –> creation of SC inventories in other languages for rendering MT evaluation more sensitive to semantics
- comparison to a French WordNet-like resource (Sagot & Fiser,'08), currently under development
- integration of the inventories in **other multilingual applications**
- exploitation of the SC inventories for **cross-lingual WSD** (Apidianaki,'09) and integration into a **Statistical MT system**

- Questions?
- This research is supported by the Science Foundation Ireland (Grant 07CEI1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.