Création rapide et efficace d'un système de désambiguïsation lexicale pour des langues peu-dotées

(Rapid construction of a supervised Word Sense Disambiguation system for under-resourced languages)

Mohammad Nasiruddin, Andon Tchechmedjiev,
Hervé Blanchon et Didier Schwab

prenom.nom@imag.fr
http://getalp.imag.fr/wsd/

LIG-GETALP
Univ. Grenoble Alpes

Outline

Introduction

Objective

Quickly build a supervised Word Sense Disambiguation (WSD) system for an under-resourced language U

Condition

Availability of

1. sense-annotated corpus from a well-resourced language W
2. Statistical Machine Translation (SMT) system $W \rightarrow U$

Proposal

Contribution

1. Script
   - to translate SemCor to language L;
     if SMT available, English $\rightarrow L$
2. Translating SemCor
   - to Bengali
   - to French
3. Building and evaluating a supervised WSD system
   - for French

Word Sense Disambiguation – 1/4

Definition

- Automatically identifying the meaning of a word from context
- A kind of classification task

Example

The mouse	is eating	cheese	.
animal	use	food	
furtively	take	flower	
manipulate	nosh	wind into	
timid person	cause	get away	
swollen bruise	worry		
electronic device	swallow		
	consume		

from Princeton WordNet 3.0

The mouse$_{\text{animal}}$ is eating$_{\text{swallow}}$ cheese$_{\text{food}}$.

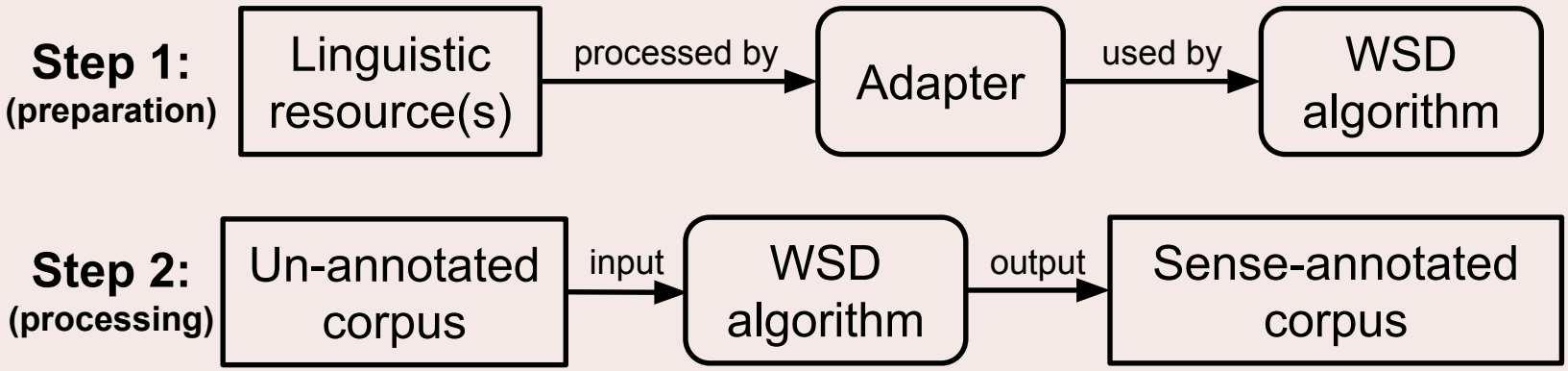
Word Sense Disambiguation – 2/4

Description

Given a context $C = w_1, ..., w_m$, to assign the correct sense(s) s_i from the set of possible senses $s_1, ..., s_n$ to some (target-words) or all words in C,

$$w_1, ..., w_m \subseteq s_1, ..., s_n$$

WSD process



Word Sense Disambiguation – 3/4

Linguistic resources exploited for WSD

1. Corpus:
   i. un-annotated – e.g.
      - Brown Corpus [Francis and Kučera, 1979]
      - British National Corpus [Burnard, 1998]
   ii. sense-annotated – e.g.
      - SemCor [Miller et al., 1993]
      - Defence Science Organization (DSO) corpus [Ng and Lee, 1996]
2. Sense inventory:
   i. less-elaborated – e.g.
      - Collins English Dictionary [Dictionary, 1979]
   ii. more-elaborated – e.g.
      - WordNet [Miller, 1995]
      - BabelNet [Navigli and Ponzetto, 2010]
      - DBnary [Sérasset, 2012]

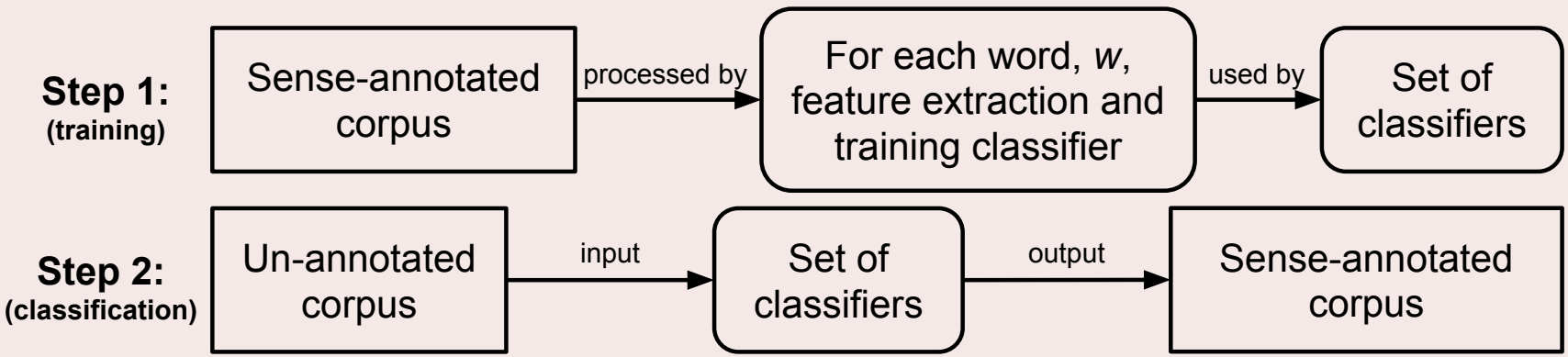
Word Sense Disambiguation – 4/4

WSD in probabilistic view

Given a context $C = w_{-m}, ..., w_{-1}[w]w_1, ..., w_n$, to assign the correct sense(s) s_i from the set of possible senses $s_1, ..., s_o$ to a word w in C,

$$\arg\max_{s_i} P(s_i|C) = \arg\max_{s_i} \frac{P(s_i)P(C|s_i)}{P(C)} = \arg\max_{s_i} P(s_i)P(C|s_i)$$

Supervised WSD process



Characterization of a language

Berment's classification [Berment, 2004]

- Measure the digitization of a language to its application – e.g
  - language script representation into memory
  - localization of different tools (e.g. word processing tool)
  - machine translation

Language named	**Score (0 – 20)**	**Resource level**	**Example**
π (Pi)	0 – 9.99	Low	Bengali, Somali, ...
μ (Mu)	10 – 14.99	Medium	Norwegian, Malay, ...
τ (Tau)	14 – 20	High	English, French, ...

✓ Considering different linguistic resources required to perform WSD and MT, Bengali is still under-resourced.

Machine Translation – 1/2

Machine Translation

A process of translating from one language to another

source language $S \rightarrow$ **MT Engine** $\rightarrow$ **target language** T

Statistical Machine Translation

- Data-driven (parallel corpus) probabilistic statistical approach
- Given a source language text S, to determine the most probable translation T in the target language,

$$\arg\max_T P(T|S) = \arg\max_T \frac{P(T)P(S|T)}{P(S)} = \arg\max_T P(T)P(S|T)$$

- Decoder
  - contains – translation model, language model, ...

Machine Translation – 2/2

Our SMT systems used for translation

1. English→French

System	Decoder	Dataset	BLEU
EN-FR [Besacier et al., 2012]	Moses	Europarl, UN, ...	24.85

2. English→Bengali

System	Decoder	*Dataset	BLEU
EN-BN	Moses	EMILLE, OPUS, INDIC, ...	27.21

***Figures**:

- Parallel corpus: 679,019 sentences – 85% for training, 10% tuning, 5% testing
- Monolingual corpus: 1,107,776 sentences (IRSTLM toolkit for Language Model)

Translating sense-annotated corpus – 1/3

State of the art of annotation transfer

Author	Application	Languages
[Brown et al., 1991]	WSD	En-Fr
[Yarowsky and Ngai, 2001]	POS-tags	En-Fr
[Yarowsky et al., 2001]	Chunks	En-Fr
[Diab and Resnik, 2002]	WSD	Fr-En
[Hwa et al., 2005]	SD	En-Cn
[Padó and Lapata, 2009]	SRL	En-De
[van der Plas and Apidianaki, 2014]	SRL	En-Fr
[Wang and Manning, 2013]	SRL	En-Fr

SD : Syntactic Dependency
SRL: Semantic Role Labeling

Translating sense-annotated corpus – 2/3

Algorithm

- *SemCor sentences (surface form, stripped SGML tags)* → Moses (shell invokation) → Translated sentences + Word alignments
  - Word alignments in an array
    - position = source word index
    - value = translated word index
  - for each instances $I_j[lemma, pos, lexsn, surfaceForm]$ indexed by j, I_j[lemma] replaced with its translation though alignment link
- Obtain a Translated SemCor

Translating sense-annotated corpus – 3/3

SemCor 3.0

```
<contextfile concordance=brown>
 <context filename=br-a01 paras=yes>
  <p pnum=1>
   <s snum=1>
    <wf cmd=ignore pos=DT>The</wf>
          <wf cmd=done rdf=group pos=NNP
lemma=group wnsn=1 lexsn=1:03:00:: pn=group>
Fulton_County_Grand_Jury</wf>
        <wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>said</wf>
    [...]
     <punc>‘‘</punc>
    <wf cmd=ignore pos=DT>no</wf>
      <wf cmd=done pos=NN lemma=evidence wnsn=1
lexsn=1:09:00::>evidence</wf>
     <punc>’’</punc>
    [...]
     <wf cmd=done pos=VB lemma=take_place wnsn=1
lexsn=2:30:00::>took_place</wf>
     <punc>.</punc>
   </s>
  </p>
  [...]
 </context>
</contextfile>
```

Translated SemCor (Bengali)

```
<contextfile concordance=brown>
 <context filename=br-a01 paras=yes>
  <p pnum=1>
   <s snum=1>
    <wf cmd=ignore pos=DT>আর</wf>
          <wf cmd=done rdf=group pos=NNP
lemma=group wnsn=1 lexsn=1:03:00:: pn=group>
Fulton_County_Grand_Jury</wf>
        <wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>বলেছিলেন</wf>
    [...]
     <punc>‘‘</punc>
    <wf cmd=ignore pos=DT>কোনো</wf>
      <wf cmd=done pos=NN lemma=evidence wnsn=1
lexsn=1:09:00::>প্রমাণ</wf>
     <punc>’’</punc>
    [...]
     <wf cmd=done pos=VB lemma=take_place wnsn=1
lexsn=2:30:00::>took_place</wf>
     <punc>।</punc>
   </s>
  </p>
  [...]
 </context>
</contextfile>
```

Supervised WSD – 1/3

Naïve Bayes classifier [Rish, 2001]

- Given a context $C = w_{-m}, ..., w_{-1}[w]w_1, ..., w_n$, from the previously training data (sense-tagged corpus) D a Baysian classifier assigns the most probable classification (sense(s)) s_i to each new instance (context) w in C.
- Naïve Bayes classifier assumes, attributes of the instance w are independent of each other,

$$\arg\max_{s_i} P(s_i)P(C|s_i) = \arg\max_{s_i} P(s_i) \prod_{w_j \in C} P(w_j|s_i)$$

Supervised WSD – 2/3

Feature set (NUS-PT system for SemEval 2007 task 7) [Chan et al., 2007]

Feature type	Number	Feature
Local collocation	11	$C_{-1,-3}$, $C_{-2,-2}$, $C_{-1,-2}$, $C_{-1,-1}$, $C_{-1,1}$, $C_{1,1}$, $C_{-1,2}$, $C_{-2,1}$, $C_{1,2}$, $C_{2,2}$, $C_{1,3}$
POS-tags	7	t_{-3}, t_{-2}, t_{-1}, t_0, t_1, t_2, t_3
Context words	depends	w_{-m}, ..., w_{-1}, w_0, w_1, ..., w_n

Example

The Fulton_County_Grand_Jury said **Friday** an investigation of Atlanta 's recent primary_election produced [...]

t_{-3}	t_{-2}	t_{-1}	t_0	t_1	t_2	t_3						
w_{-3}	w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	[...]

$c_{-1,-3}$: The Fulton_County_Grand_Jury said
$c_{-2,-2}$: Fulton_County_Grand_Jury
$c_{-1,-2}$: Fulton_County_Grand_Jury said
$c_{-1,-1}$: said
$c_{-1,1}$: said an
$c_{1,1}$: an
$c_{-1,2}$: said an investigation
$c_{-2,1}$: Fulton_County_Grand_Jury said an
$c_{1,2}$: an investigation
$c_{2,2}$: investigation
$c_{1,3}$: an investigation of

Supervised WSD – 3/3

Three systems

1. **For English**
   - directly trained with SemCor 3.0
2. **For Bengali**
   - trained with translated SemCor 3.0 with transferred annotations (Bengali)
3. **For French**
   - trained with translated SemCor 3.0 with transferred annotations (French)

Evaluation – 1/4

Evaluation corpus (SemEval 2013 task 12 multilingual WSD)[Navigli et al., 2013]

- **5 languages** : de, en, es, fr, it
- **Corpus** : 13 articles/language (different domains)
- **WSD task** : all-nouns
- **Sense inventory**: BabelNet 1.1.1 (WordNet 3.0 + Wikipedia)

Challenge

- Sense inventory for SemCor 3.0 – Princeton WordNet 3.0
- Sense mapping between – Princeton WordNet 3.0 and BabelNet 1.1.1
- Mapping coverage: **96.10%**

Evaluation – 2/4

Evaluation measures

1. *Precision P*
   - number of correct senses among the number of senses returned by the system
2. *Recall R*
   - number of correct senses among the number of expected senses from the system
3. *F1-score*
   - harmonic mean of *Precision* and *Recall*, $\frac{2.P.R}{P+R}$

Evaluation – 3/4

Result

Trained on	*System*	*P*	*R*	*F1*
SemCor	SUP-EN	64.80%	64.70%	64.75%
	MFS-EN	66.90%	66.60%	66.64%
	difference	*-2.10%*	*-1.90%*	*-1.89%*
Translated SemCor*	SUP-FR	51.60%	51.50%	51.55%
	MFS-FR	45.60%	45.10%	45.34%
	difference	**+6.00%**	**+6.40%**	**+6.21%**

*Transferred annotations

SUP-EN: English supervised WSD system
MFS-EN: English Most Frequence Sense baseline
SUP-FR: French supervised WSD system
MFS-FR: French Most Frequence Sense baseline

Evaluation – 4/4

Comparison

Language	*System*	*P*	*R*	*F1*
EN	UMCC-DLSI**	68.50%	68.50%	68.50%
	GETALP-SUP	**64.80%**	**64.70%**	**64.75%**
	DAEBAK!**	60.40%	60.40%	60.40%
	GETALP-UNS*	58.30%	58.30%	58.30%
FR	UMCC-DLSI**	60.50%	60.50%	60.50%
	DAEBAK!**	53.80%	53.80%	53.80%
	GETALP-SUP	**51.60%**	**51.50%**	**51.55%**
	GETALP-UNS*	48.30%	48.20%	48.30%

*[Schwab et al., 2013]
**[Navigli et al., 2013]

GETALP-SUP: GETALP supervised WSD system
GETALP-UNS: GETALP unsupervised WSD system
DAEBAK! : DAEBAK! unsupervised WSD system
UMCC-DLSI : UMCC-DLSI unsupervised WSD system

Conclusion and perspective

Conclusion

- Translating a sense-annotated corpora and transferring sense-annotations though SMT word alignments
- Quickly built a supervised WSD system for any language
- Acceptable disambiguation performance with no tuning and little training data

Perspective

- Training the system with more annotated data (e.g. DSO)
- Compare different supervised algorithms and classifier fusion techniques (e.g. Boosting [Zhou, 2012])
- Combine supervised and unsupervised approaches

https://getalp.imag.fr/wsd/

Thank you!

Questions?

Reference I

Berment, V. (2004).
Méthodes pour informatiser les langues et les groupes de langues «peu dotées».
PhD thesis, Université Joseph-Fourier-Grenoble I.

Besacier, L., Lecouteux, B., Azouzi, M., and Luong, N.-Q. (2012).
The lig english to french machine translation system for iwslt 2012.
In *IWSLT*, pages 102–108.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1991).
Word-sense disambiguation using statistical methods.
In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 264–270. Association for Computational Linguistics.

Burnard, L. (1998).
The British National Corpus.

Chan, Y. S., Ng, H. T., and Zhong, Z. (2007).
Nus-pt: exploiting parallel texts for word sense disambiguation in the english all-words tasks.
In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 253–256. Association for Computational Linguistics.

Diab, M. and Resnik, P. (2002).
An unsupervised method for word sense tagging using parallel corpora.
In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262. Association for Computational Linguistics.

Reference II

Dictionary, C. E. (1979).
Collins.

Francis, W. N. and Kučera, H. (1979).
Brown corpus manual.
Brown University.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005).
Bootstrapping parsers via syntactic projection across parallel texts.
Natural language engineering, 11(03):311–325.

Miller, G. A. (1995).
Wordnet: a lexical database for english.
Communications of the ACM, 38(11):39–41.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993).
A semantic concordance.
In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Navigli, R., Jurgens, D., and Vannella, D. (2013).
Semeval-2013 task 12: Multilingual word sense disambiguation.
In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 222–231.

Reference III

Navigli, R. and Ponzetto, S. P. (2010).

Babelnet: Building a very large multilingual semantic network.
In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Ng, H. T. and Lee, H. B. (1996).

Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach.
In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2009).

Cross-lingual annotation projection for semantic roles.
Journal of Artificial Intelligence Research, 36(1):307–340.

Rish, I. (2001).

An empirical study of the naive bayes classifier.
In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York.

Schwab, D., Tchechmedjiev, A., Goulian, J., Nasiruddin, M., Sérasset, G., and Blanchon, H. (2013).

Getalp system: Propagation of a lesk measure through an ant colony algorithm.
In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 232–240. Association for Computational Linguistics.

Reference IV

Sérasset, G. (2012).
Dbnary: Wiktionary as a lmf based multilingual rdf network.
In *Language Resources and Evaluation Conference, LREC 2012*.

van der Plas, L. and Apidianaki, M. (2014).
Cross-lingual word sense disambiguation for predicate labelling of french.
Articles longs, page 46.

Wang, M. and Manning, C. D. (2013).
Cross-lingual pseudo-projected expectation regularization for weakly supervised learning.
arXiv preprint arXiv:1310.1597.

Yarowsky, D. and Ngai, G. (2001).
Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora.
In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001).
Inducing multilingual text analysis tools via robust projection across aligned corpora.
In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Zhou, Z.-H. (2012).
Ensemble methods: foundations and algorithms.
CRC Press.