WSD for n-best reranking and local language modeling in SMT

Marianna Apidianaki, Guillaume Wisniewski,
Artem Sokolov, Aurélien Max, François Yvon

LIMSI-CNRS & Univ. Paris Sud
Orsay, France

Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)
Jeju, Korea, 12 July 2012

Towards integrating some semantics into SMT

Some open issues in WSD for SMT

- type of context used for disambiguation
- types of disambiguated words
- disambiguated units
- single classifier vs unit-dependent classifier
- type of integration for the WSD predictions

Towards integrating some semantics into SMT

Some open issues in WSD for SMT

- type of context used for disambiguation
- types of disambiguated words
- disambiguated units
- single classifier vs unit-dependent classifier
- type of integration for the WSD predictions

This work is a preliminary attempt that

- disambiguates content words only
- disambiguates at the level of individual forms
- experiments with two methods for integrating the predictions
- reports contrastive results w.r.t. a baseline system

Outline

Task-oriented multilingual WSD

Word Sense Disambiguation (WSD)

task of identifying the sense of words in texts

Task-oriented WSD

- aims to improve the performance of complex NLP systems (Ide and Wilks, 2007)
  - **unsupervised methods** oriented towards the disambiguation needs of multilingual applications
  - use of **senses** relevant to multilingual applications identified by the **translations of words or phrases** in a parallel corpus (Carpuat and Wu, 2007; Chan et al, 2007) or by **more complex representations** generated by word sense induction methods (Apidianaki, 2009)

Related work

- **Carpuat and Wu (2005)** integrate WSD predictions into a SMT system
  1. constrain the set of translations considered by the decoder for each target word
  2. replace the translation of each target word by the WSD prediction
- **Carpuat and Wu (2007), Stroppa et al. (2007)** generalize a WSD system so that it performs fully phrasal multiword disambiguation
- **Chan et al. (2007)** modify the rule weights of a hierarchical translation system to reflect the predictions of their WSD system
- **Haque et al. (2009)** and **(2010)** introduce lexico-syntactic descriptions in the form of supertags as source language context-informed features in a PB-SMT and a hierarchical model
- **Mauser et al. (2009)** and **Patry and Langlais (2011)** train a global lexicon model that predicts the bag of output words from the bag of input words

Towards integrating semantics into SMT

Objective of this work

Investigate the **impact** of integrating the **predictions of a cross-lingual WSD classifier** into an SMT system in two ways :

1. by **reranking** the translations in the n-**best list** generated by the SMT system
2. by a tighter integration of the WSD classifier with the rest of the system by estimating an additional ***sentence specific*** **language model** that exploits the WSD predictions and is used during decoding

Outline

The WSD classifier

Variation of the classifier proposed in (Apidianaki, 2009)

- contextual disambiguation of words by selecting the most appropriate cluster of translations
- candidate clusters (semantically similar translations) are built by a cross-lingual word sense induction method
- here, the classifier simply discriminates between unclustered translations of a word and assigns a score to each translation for each disambiguated word instance
- **translations** are represented by a **source language feature vector** that the classifier uses for disambiguation

Data and preprocessing

- Use of the **TED talks EN-FR training data** (from IWSLT'11)
  - 107,268 parallel sentences
  - word alignment in both directions using GIZA++
- **Bilingual lexicons** are built from the resulting alignments which are filtered to eliminate spurious alignments
  - translations with a probability lower than 0.01 are discarded
  - translations are filtered by PoS
  - only intersecting alignments are kept
  - lexicon entries that have more than 20 translations after filtering are not considered

Vector building

A **vector** is built for **each translation** T_i of an EN word w

- the **features** of the vector of a T_i are the **lemmas** of the **content words** that co-occur with w in the corresponding **source sentences** of the parallel corpus
- **each feature** F_j (1< j<N) receives a **total weight** with a T_i

Total weight

$$\mathrm{tw}(F_j, T_i) = \mathrm{gw}(F_j) \cdot \mathrm{lw}(F_j, T_i) \tag{1}$$

Global weight

$$\mathrm{gw}(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{ij})}{N_i} \quad (2)$$

- N_i: the number of translations (T_i's) to which F_j is related
- p_{ij}: the probability that F_j co-occurs with instances of w translated by T_i

$$p_{ij} = \frac{\mathrm{cooc\_frequency}(F_j, T_i)}{N} \quad (3)$$

- $\mathrm{cooc\_frequency}(F_j, T_i)$: co-occurrence frequency of F_j with w when translated as T_i
- N: total number of features seen with T_i

Local weight

$$\mathrm{lw}(F_j, T_i) = \log(\mathrm{cooc\_frequency}(F_j, T_i)) \quad (4)$$

The WSD classifier

- Vectors contain **lemmas** but we disambiguate word forms
- WSD is performed by comparing
  - the vector associated with each translation T_i of a word w
  - the context of each occurrence of w in the input sentences
- A (normalized) score for each translation of each occurrence of w is returned :

$$\text{assoc\_score}(V_i, C) = \frac{\sum_{j=1}^{|CF|} \text{tw}(CF_j, T_i)}{|CF|} \tag{5}$$

- $(CF_j)_{j=1}^{|CF|}$: the set of common features between vector V_i and context C
- tw: the weight of a CF_j with translation T_i

The WSD classifier : example

you know, one of the intense_{intenses (0.305), forte (0.306), intense (0.389)} pleasures of travel_{transport (0.334), voyage (0.332), voyager (0.334)} and one of the delights of ethnographic research_{recherche (0.225), research (0.167), études (0.218), recherches (0.222), étude (0.167)} is the opportunity_{possibilité (0.187), chance (0.185), opportunités (0.199), occasion (0.222), opportunité (0.207)} to live amongst those who have not forgotten_{oublié (0.401), oubliés (0.279), oubliée (0.321)} the old_{ancien (0.079), âge (0.089), anciennes (0.072), âgées (0.100), âgés (0.063), ancienne (0.072), vieille (0.093), ans (0.088), vieux (0.086), vieil (0.078), anciens (0.081), vieilles (0.099)} ways_{façons (0.162), manières (0.140), moyens (0.161), aspects (0.113), façon (0.139), moyen (0.124), manière (0.161)}, who still feel their past_{passée (0.269), autrefois (0.350), passé (0.381)} in the wind_{éolienne (0.305), vent (0.392), éoliennes (0.304)}, touch_{touchent (0.236), touchez (0.235), touche (0.235), toucher (0.293)} it in stones_{pierres(1.000)} polished by rain_{pluie (1.000)}, taste_{goût(0.500), goûter(0.500)} it in the bitter_{amer (0.360), amère (0.280), amertume (0.360)} leaves_{feuilles (0.500), feuillages (0.500)} of plants_{usines (0.239), centrales (0.207), plantes (0.347), végétaux (0.207)}.

Coverage of the WSD method

PoS	# of words	# of WSD predictions	%
Nouns	5535	3472	62.72
Verbs	5336	1269	23.78
Adjs	1787	1249	69.89
Advs	2224	1098	49.37
all content PoS	14882	7088	47.62

- Focus on prediction with higher confidence
- For instance, only 1/4 of English verbs are disambiguated

Outline

n-best reranking

- Simple way to bias hypothesis selection with WSD
  - avoids tight integration with decoder
  - limited to hypotheses that survived pruning
- Add feature(s) to reflect WSD variants' usage rate in hypotheses
  - `wsd-sum`: add probabilities of matching translation variants
  - `wsd-norm-sum`: `wsd-sum` divided by the number of source words

`src`: intense$_{\{intenses(0.305), forte(0.306), intense(0.389)\}}$ pleasures of travel$_{\{transport(0.334), voyage(0.332), voyager(0.334)\}}$

$\mathtt{hyp}_1$:	immense plaisir de metro	`wsd-sum`: 0.000, `wsd-norm-sum`: 0.000
$\mathtt{hyp}_2$:	plaisir forte de voyages	`wsd-sum`: 0.306, `wsd-norm-sum`: 0.076
$\mathtt{hyp}_3$:	plaisirs intenses de voyage	`wsd-sum`: 0.637, `wsd-norm-sum`: 0.159

- Rerun MERT on augmented *n*-best lists to get new model weights

Local language models

- Use an additional language model to directly integrate the prediction of the WSD system into the decoder (Crego et al., 2010)
  1. for each source sentence, estimate an additional language model
  2. use this language model during decoding
- Each translation predicted by the WSD classifier can be scored by the additional LM
  - use the probability of the WSD classifier
  - use a small arbitrary constant for "unknown" words
- Several advantages
  - no hard decisions are made when integrating WSD predictions
  - disambiguation is automatically propagated at the phrase level
  - WSD predictions are applied before search space pruning

Outline

Experimental setting : data

- TED-talk English to French dataset provided by the IWSLT'11 evaluation campaign
  - a monolingual corpus (111,431 sentences) used to estimate a 4-gram language model with KN-smoothing
  - a bilingual corpus (107,268 sentences) used to extract the phrase table
  - all data tokenized, cleaned and lowercased
  - English side PoS-tagged with `TreeTagger` (Schmid, 1994)
- System optimizations using MERT on dev-2010 (934 sentences)
- Evaluations performed on test-2010 (1,664 sentences)

Experimental setting : baseline systems

- Standard PBSMT decoder `Moses` (Koehn et al., 2007) with standard training pipeline
  - bitext alignment using GIZA++, symmetrization, `grow-diag-final-and` heuristic, bi-phrase extraction and scoring
- Use the **IBM 1 model** estimated during the SMT system training as a (naive) WSD system
  - one of the **best performing features** for *n*-best list reranking (Och et al., 2004)
  - define a sentence-level **additional language model** with the 20 best translations according to the IBM 1 model and their probability

Experimental setting : oracle systems

- Run **oracle** experiments of (Crego et al., 2010) to estimate an upper bound on performance
  - train of a sentence-level language model using the reference translation
  - amounts to assuming that the WSD system correctly predicted a single translation for each word
  - however, for that experiment all source words (i.e. the whole reference translation) were "disambiguated"

Experiments : results

method		BLEU	METEOR
baseline	—	29.63	53.78
rescoring	WSD (zero init)	30.00	54.26
	WSD (no reinit)	29.58	53.96
additional LM	oracle 1-gram	42.92	69.39
	IBM 1	30.18	54.36
	WSD	30.51	54.38

- baseline $<$ rescoring $<$ additional LM
- In additional LM, WSD only improves over IBM 1 on BLEU (+0.33) (score used for tuning)
- The oracle shows important room for improvement, but recall:
  - that we disambiguate at the form level
  - that we used a single reference translation
  - that all source words were "disambiguated" by the oracle (can have some negative impact)

Results : contrastive evaluation

- Fine-grained evaluation of the translations produced by different systems on word classes for the source language (Max et al., 2010)
  - compare how source words are translated by two systems: the `Moses` baseline and our WSD-informed system using additional LM
  - use PoS for content words as source classes
  - count a word as correctly translated when its translation belongs to the reference translation (report percentage)

PoS	baseline	WSD	Δ
Nouns	67.57	69.06	+1.49
Verbs	45.97	47.76	+1.79
Adjectives	51.79	53.94	+2.15
Adverbs	52.17	56.25	+4.08

- Best (absolute) performance on Nouns
- All PoS improved, highest improvement on Adverbs

Results : contrastive evaluation of context words

- Study whether word translation disambiguation influences the translation of surrounding words

	baseline				WSD			
	w_{-2}	w_{-1}	w_{+1}	w_{+2}	w_{-2}	w_{-1}	w_{+1}	w_{+2}
Nouns	64.0	68.6	75.2	64.6	65.5	70.5	76.3	66.6
Verbs	68.6	67.5	63.0	62.2	70.0	68.9	64.8	64.2
Adjectives	63.1	64.4	64.3	66.5	64.1	65.6	64.8	69.3
Adverbs	70.8	69.4	68.7	66.4	71.0	71.2	70.0	67.2

- Positive impact of WSD on the translation of surrounding words
- Note : some context words from the immediate context may have been directly (correctly or incorrectly) disambiguated

Outline

Conclusions

- Preliminary study on **WSD prediction integration into SMT**
  - treats only single words (no phrases)
  - restrictive definition of disambiguated words (only 47% of CWs)
  - predicts at the form level (no target-side sense clusters)
- Encouraging results
  - both n-**best list rescoring** and **local language model approaches** can successfully exploit the WSD predictions
  - the contrastive evaluation shows that surrounding (target) words also benefit from these improvements
  - the initial oracle study shows that there is still room for improvement (although it cannot be attributed entirely to WSD predictions)

Future work

- Use of **translation sense clusters** (Apidianaki, 2009; Bansal et al., 2012)
  - for improving **MT lexical choice**
  - for **semantics-sensitive MT evaluation**
- **Disambiguation** at the level of **lemmas**
  - sparseness reduction
  - handling lemmatized predictions in SMT
- **Extension of the coverage** of the WSD method
  - disambiguation of phrases

WSD for n-best reranking and local language modeling in SMT

Marianna Apidianaki, Guillaume Wisniewski,
Artem Sokolov, Aurélien Max, François Yvon

LIMSI-CNRS & Univ. Paris Sud
Orsay, France

Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)
Jeju, Korea, 12 July 2012