

Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-Vocabulary Words

Wilker Aziz⁽¹⁾, Marc Dymetman⁽²⁾, Shachar Mirkin⁽³⁾,
Lucia Specia⁽⁴⁾, Nicola Cancedda⁽²⁾ and Ido Dagan⁽³⁾

(1) São Paulo University (2) Xerox Research Centre Europe;
(3) Bar Ilan University (4) University of Wolverhampton

OOV words and paraphrase/entailment

The mayor was *attacked* by the press

Le maire a été *attacked* par la presse

phrase pairs
(aka biphrases)

(**mayor**, **maire**)

(**press**, **presse**)

(**attacked**, ?)



OOV words and paraphrase/entailment

The mayor was *attacked* by the press

Le maire a été *attacked* par la presse

phrase pairs
(aka biphrases)

(**mayor**, maire)

(**press**, presse)

(**attacked**, ?)



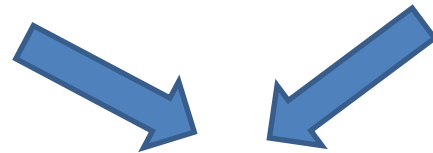
entailments

attacked → **accused**

attacked → **hit**

(**accused**, accusé)

(**hit**, touché)



(**attacked**, accusé)

(**attacked**, touché)

Le maire a été accusé par la presse

Callison-Burch et al, 2006; Marton et al, 2009: **paraphrases**

Mirkin et al, 2009: **entailments**

OOV entailments as a learning problem

- By contrast to previous work on paraphrase/entailments for OOV in SMT, we cast replacement selection as:
 - a learning problem
 - from human annotations
 - with the entailment model tightly integrated into the Phrase-Based SMT decoder
- **Learning an OOV expert for a PB-SMT system**

Learning an expert for OOV sentences

- **Integrated PB-SMT model, that includes an expert for OOV sentences**
 - One overall SMT model, built on top of standard PB-SMT model
 - Contextual features, representing properties of the replacements
 - “Dynamic” biphases built on demand
- **Learning from human judgments**
 - Annotators rank translations corresponding to different replacement choices
 - The integrated SMT model is tuned in order to bring system ranking close to annotator ranking
- **Active learning:**
 - For each OOV sentence, only a few candidate translations are shown, depending on current state of the model
- **Avoiding to bias the SMT system towards OOV sentences**
 - Learning is done in such a way that the integrated model behaves like the standard model on standard sentences

Dynamic biphrases


Source
entailment

attacked → accused
attacked → hit

Static
biphrase

(accused, accusé)

(hit, touché)



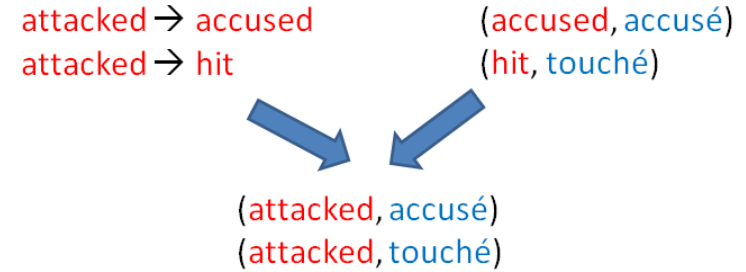
(attacked, accusé)
(attacked, touché)

Dynamic
biphrase

Features

Entailment features

	<i>DSim</i>	<i>CSim</i>	<i>InfoLoss</i>
attacked → accused	-3.1	-0.3	-0.4
attacked → hit	-5.2	-7.2	-0.5



Biphrase features

	source	target	static features		dynamic features			
			<i>For</i>	<i>Rev</i>	<i>DSim</i>	<i>CSim</i>	<i>InfoLoss</i>	<i>Clone</i>
static biphrases	mayor	maire	-0.1	-0.1	0	0	0	0
	press	presse	-1.5	-0.7	0	0	0	0
	accused	accusé	-1.6	-1.2	0	0	0	0
	hit	touché	-0.9	-0.5	0	0	0	0
dynamic biphrases	attacked	accusé	0	0	-3.1	-0.3	-0.4	-1.6
	attacked	touché	0	0	-5.2	-7.2	-0.5	-0.9

Entailment features: details

- Contextual score (CSIM):
 - How well does *rep* fit the context of the sentence *s*
 - Based on cosine similarity of LSA vectors representing *rep* and *s*
- Domain similarity (DSIM):
 - How well does *rep* replaces *oov* in general texts of the domain
 - Based on cosine similarity of LSA vectors of *oov* and *rep*
- Information Loss (InfoLoss):
 - Measures distance between *oov* and *rep* in Wordnet
- Other entailment features:
 - Synonym/Hypernym (from Wordnet)
 - Identity replacement (replacement by copying source word)

The integrated model

- Original model

G: standard “static” features

$$\operatorname{argmax}_{(a,t)} \Lambda \cdot G(s,t,a)$$

- Integrated model

H: “dynamic” features

$$\operatorname{argmax}_{(a,t)} \Lambda \cdot G(s,t,a) + M \cdot H(s,t,a)$$

- Integrated feature vector: $F = G \oplus H$
- Integrated parameter vector: $\Omega = \Lambda \oplus M$

Human annotations:

(1) Active sampling

Given an initial value for $\Omega = \Lambda \oplus M$, and an OOV sentence s ...

... actively sample around a dozen different translations for s (out of many more candidates)

- According to probabilities assigned by Ω to these translations (but always include Ω -best translation)
- But also including top candidates relative to individual features (contextual score, domain similarity, ...)

Human annotations:

(2) Annotation interface

- Present these translations in an annotation interface
 - Ask annotators to concentrate on “closeness of meaning” for portions affected by the replacements
 - BLEU would be inadequate for this
 - Discourage too fine distinctions:
 - translations grouped in a few clusters

Human annotations:

(3) Update the parameters

- Update Ω from the annotation data:
 - Try to bring model rank and annotation rank closer
 - Whenever two translations (s,t_j) and (s,t_k) are ordered differently by the annotator and by the model
- ... then change Ω into Ω' , in such a way that:
1. Ω' now ranks (s,t_j) and (s,t_k) in the same order as the annotator
 2. Ω' moves from Ω as little as possible (in terms of Euclidian distance)
 3. If $\Omega = \Lambda \oplus M$, then $\Omega' = \Lambda \oplus M'$ (update does not change Λ)

Adaptation
of MIRA

Model preserves
behavior on non OOV
sentences

Human annotations:

(3) Update the parameters

- Use Ω' for the next round of active sampling
 - For efficiency, Ω is only updated after batches of 80 source sentences

Experimental setup

- Baseline phrase-based SMT system: MATRAX
 - Trained on English-French Europarl data (1M sents)
- Training of integrated expert model:
 - 75,000 sents from WMT-09 News Commentary
 - Around 15% OOV sentences
 - Tuning set: 1,000 OOV sents
 - Two annotators
 - Active sampling on batches of 80 sents
 - Convergence of performance after 6 slices (480 sents)
 - Evaluation set: 500 OOV sents
 - Comparison of different systems

Results

System	μ	σ	Best	Acceptance
Expert-Human'	2.274	1.803	0.6258	0.7002
Mirkin09-1	2.736	1.933	0.5172	0.5822
Mirkin09-2	2.744	1.931	0.5132	0.5822
Expert-Human	3.018	1.913	0.4145	0.5252
Expert-MERT	3.153	1.928	0.4024	0.4849
SMT-Baseline	3.998	1.603	0.1549	0.2918
Stat-Paraphrases	4.107	1.584	0.1690	0.2495

- *SMT-baseline*: The base SMT system MATRAX
- *Mirkin09-1,-2*: Two best 'entailment' systems from Mirkin et al, 2009: replacement choices not integrated in decoder, no training of the expert
- *Stat-Paraphrases*: An implementation of Marton-09, with new static biphases obtained through paraphrases from original static biphases
- ***Expert-Human***: The model of this paper, trained from human annotations
- ***Expert-Human'***: Identity replacements blocked at decoding time
- *Expert-Mert*: The model of this paper but trained by MERT

Conclusion

- OOV: an instance of a more general problem: **Learning an Expert for SMT**

On the basis of an existing SMT system

... And on a narrow domain of « expertise »

... Improve the performance of the system

... Based on human judgments for the narrow domain

... Without degrading the behavior of the system on sentences outside of the narrow domain

BACKUP

Thanks!

Learning to rank using MIRA

If $\Omega \cdot \Phi(y_{j,k}) \geq 0$ then $\Omega' := \Omega$

Else $\Omega' := \operatorname{argmin}_{\omega} \|\omega - \Omega\|^2$

s.t. $\omega \cdot \Phi(y_{j,k}) - \omega \cdot \Phi(y_{k,j}) \geq 1$

and $\pi^{(\Lambda)}(\omega) = \Lambda_0$

Feature combinations

Features	μ	σ	Best	Acceptance
LID	2.477	1.465	0.4728	0.5252
ID	2.491	1.463	0.4668	0.5211
LI	2.547	1.457	0.4427	0.5050
I	2.561	1.463	0.4447	0.4970
D	2.924	1.414	0.3360	0.3722
LD	2.930	1.412	0.3340	0.3702
L	3.056	1.361	0.2857	0.3300
Baseline	3.219	1.252	0.2093	0.2918

Table 1: Comparison between different feature combinations and the baseline showing the percentage of times each combination outputs a translation that is acceptable, i.e. is not discarded (Acceptance), a translation that is ranked in the first cluster (Best), as well as the the mean rank (μ) and standard deviation (σ) of each combination, where the discarded translations are conventionally assigned a rank of 5, lower than the rank of any acceptable cluster observed among the annotations. (L) context model score, (I) information-loss, (D) domain similarity, (Baseline) SMT system.

Learning iterations improve results

Iterations	μ	σ	Best	Acceptance
M_6	2.487	1.458	0.4628	0.5252
M_5	2.491	1.459	0.4628	0.5231
M_4	2.489	1.458	0.4628	0.5252
M_3	2.493	1.455	0.4588	0.5252
M_2	2.501	1.456	0.4567	0.5211
M_1	2.519	1.456	0.4507	0.5151
M_0	2.944	1.407	0.328	0.3642
Baseline	3.237	1.228	0.1932	0.2918