Learning Morphological Normalization for Translation from and into Morphologically Rich Languages

Franck Burlot, François Yvon

May 29, 2017

EAMT, Prague, Czech Republic

Introduction

Target morphology difficulties

- Dissymmetry of both languages involved is hard to handle:

English	I will go by car.	Jan loves Hana.
Czech	pojedu autem.	Hanu miluje Jan.

- One English word can translate into several Czech words:

English	**Czech**
beautiful	krásný krásného krásnému krásném krásným krásná krásné krásnou krásní krásných krásnými

- Many sparsity issues (OOVs)
- The translation probability of a Czech word form is hard to estimate when its frequency is low in the training data.

➡ **Idea**: Simplify the translation process by making Czech look like English (beautiful $\rightarrow$ krásn$\emptyset$).

➡ **Assumption**: Such a simplification could make translation easier from and into the morphologically rich language (MRL).

A Clustering Algorithm

Clustering the source-side MRL

- Goal: cluster together MRL forms that translate into the same target word(s).
- Words are represented as a lemma and a fine-grained PoS: autem $\rightarrow$ auto+Noun+Neut+Sing+Inst
- We have one lemma and $\mathbf{f}$, all the word forms in its paradigm.
- $\mathbf{E}$ is the complete English vocabulary.

Conditional entropy of the translation model

$$
\begin{aligned}
H(\mathbf{E}|\mathbf{f}) &= \sum_{f \in \mathbf{f}} p(f) H(\mathbf{E}|f) \\
&= \sum_{f \in \mathbf{f}} \frac{p(f)}{\log_2 |\mathbf{E}_{a_f}|} \sum_{e \in \mathbf{E}_{a_f}} p(e|f) \log_2 p(e|f)
\end{aligned}
$$

Information Gain (IG)

- Start with an initial state where each form in $\mathbf{f}$ is a singleton cluster.
- Repeatedly try to merge cluster pairs (f_1 and f_2) so as to reduce the conditional entropy.
- f' is the resulting cluster from the merge.

Compute IG for every cluster pairs

$$\begin{aligned}\mathrm{IG}(f_1, f_2) = {} & p(f_1)H(\mathbf{E}|f_1) \\ & + p(f_2)H(\mathbf{E}|f_2) \\ & - p(f')H(\mathbf{E}|f')\end{aligned}$$

Source-side Clustering

- In practice, the algorithm is applied at the level of PoS, rather than individual lemmas.
- For a given PoS, all lemmas have the same number of possible morphological variants (cells in their paradigm).
- Our goal is to cluster the paradigm cells.
- Since we can't set the optimal number of clusters in advance, we opted for an agglomerative clustering procedure.

Initial State

- Input to the algorithm:

Word Form	Unigram	Alignments	Entropy
kočka+Noun+Sing+Nominative	0.01	cat (0.9), kitten (0.1)	0.47
kočka+Noun+Sing+Accusative	0.02	cat (0.8), kitten (0.2)	0.72
pes+Noun+Sing+Nominative	0.05	dog (0.95), puppy (0.05)	0.29
pes+Noun+Sing+Accusative	0.03	dog (0.9), puppy (0.1)	0.47
kočka+Noun+Plur+Nominative	0.09	cats (0.8), kittens (0.15), cat (0.005)	0.56
pes+Noun+Plur+Nominative	0.09	dogs (0.9), puppies (0.08), dog (0.002)	0.28

Initial State

- Input to the algorithm:

Word Form	Unigram	Alignments	Entropy
kočka+Noun+Sing+Nominative	0.01	cat (0.9), kitten (0.1)	0.47
kočka+Noun+Sing+Accusative	0.02	cat (0.8), kitten (0.2)	0.72
pes+Noun+Sing+Nominative	0.05	dog (0.95), puppy (0.05)	0.29
pes+Noun+Sing+Accusative	0.03	dog (0.9), puppy (0.1)	0.47
kočka+Noun+Plur+Nominative	0.09	cats (0.8), kittens (0.15), cat (0.005)	0.56
pes+Noun+Plur+Nominative	0.09	dogs (0.9), puppies (0.08), dog (0.002)	0.28

- When we start, each cluster contains a singleton word form:

Word Form	Unigram	Alignments	Entropy
kočka+Noun+0	0.01	cat (0.9), kitten (0.1)	0.47
kočka+Noun+1	0.02	cat (0.8), kitten (0.2)	0.72
pes+Noun+0	0.05	dog (0.95), puppy (0.05)	0.29
pes+Noun+1	0.03	dog (0.9), puppy (0.1)	0.47
kočka+Noun+2	0.09	cats (0.8), kittens (0.15), cat (0.005)	0.56
pes+Noun+2	0.09	dogs (0.9), puppies (0.08), dog (0.002)	0.28

- Where Noun+0 = {Sing+Nominative}

Lemma-level IG Matrices

- Compute the IG obtained for merging kočka+Noun+0 and kočka+Noun+1:

$$\begin{aligned} IG(\text{kočka+Noun+0}, \text{kočka+Noun+1}) = {} & p(\text{kočka+Noun+0}) H(\mathbf{E} | \text{kočka+Noun+0}) \\ & + p(\text{kočka+Noun+1}) H(\mathbf{E} | \text{kočka+Noun+1}) \\ & - p(\text{kočka+Noun+0:1}) H(\mathbf{E} | \text{kočka+Noun+0:1}) \end{aligned}$$

Lemma-level IG Matrices

- Compute the IG obtained for merging kočka+Noun+0 and kočka+Noun+1:

$$
\begin{aligned}
IG(\text{kočka+Noun+0}, \text{kočka+Noun+1}) = {} & p(\text{kočka+Noun+0}) H(\mathbf{E} | \text{kočka+Noun+0}) \\
& + p(\text{kočka+Noun+1}) H(\mathbf{E} | \text{kočka+Noun+1}) \\
& - p(\text{kočka+Noun+0:1}) H(\mathbf{E} | \text{kočka+Noun+0:1})
\end{aligned}
$$

- Repeat for every pairs of clusters to obtain the lemma-level IG Matrix for *kočka*:

	0	**1**	**2**
0		0.0008	-0.022
1	0.0008		-0.027
2	-0.022	-0.027	

Pos-level Matrices

- All lemma-level matrices are combined in order to get a PoS-level matrix M.
- We introduce two ways to obtain M.

PoS-level Matrices: method 1

- Sum over all the lemma-level matrices to obtain the PoS-level matrix M:

kočka

	0	1	2
0		0.0008	-0.022
1	0.0008		-0.027
2	-0.022	-0.027	

+

pes

	0	1	2
0		0.0024	-0.085
1	0.0024		-0.071
2	-0.085	-0.071	

=

Noun

	0	1	2
0		0.0032	-0.107
1	0.0032		-0.098
2	-0.107	-0.098	

PoS-level Matrices: method 2

M can be treated like a similarity matrix and updated using a procedure reminiscient of the linkage clustering algorithm:

$$M(c_1, c_2) = \frac{\sum_{f_1 \in c_1} \sum_{f_2 \in c_2} M(f_1, f_2)}{|c_1| \times |c_2|}$$

➡ This second method gives a better runtime with nearly no impact on the produced clustering. (see experimental results)

Merge

Noun

	0	1	2
0		0.0032	-0.107
1	0.0032		-0.098
2	-0.107	-0.098	

- Get the argmax from PoS-level matrix M:

$$\arg\max_{i,j} M(i,j) = 0, 1$$

- Does $M[0,1]$ exceed the threshold value $m = 0$?

Merge

Noun

	0	1	2
0		0.0032	-0.107
1	0.0032		-0.098
2	-0.107	-0.098	

- Get the argmax from PoS-level matrix M:

$$\arg\max_{i,j} M(i,j) = 0, 1$$

- Does $M[0,1]$ exceed the threshold value $m = 0$? **YES**
- Merge Noun+0 and Noun+1 in the initial set of clusters.
- New set of clusters for PoS Noun: {Noun+0, Noun+1}

Repeat with the new set of clusters

- As a result, we obtain the new PoS-level matrix M:

Noun

	0	1
0		-0.109
1	-0.109	

- Get the argmax:

$$\arg\max_{i,j} M(i,j) = 0, 1$$

- Since $M[0,1]$ does not exceed $m = 0$, the procedure stops.

Result of the procedure

In the end, we obtain the following clustering of noun paradigms, that can be applied to the MRL in different ways:

- **Cluster Noun+0**: {Sing+Nominative, Sing+Accusative}
- **Cluster Noun+1**: {Plur+Nominative}

In Practice

- Alignments used to train normalization are learnt with Fastalign.
- Filter out lemmas appearing less than 100 times and word forms with a frequency lower than 10.
- We set the minimum IG for a merge to 0.

Experiments

Setup

- Moses systems
- 4-gram LMs with KenLM
- Datasets:

	cs2en		en2cs		cs2fr		ru2en	
Setup	parall	mono	parall	mono	parall	mono	parall	mono
`Small`	190k	150M	190k	8.4M	622k	12.3M	190k	150M
`Larger`	1M	150M	1M	34.4M				
`Largest`	7M	250M	7M	54M				

- MRL clustering is performed independently for each dataset (except `Larger` and `Largest` Czech systems trained on `Larger`).
- Czech PoS obtained with Morphodita
- Russian PoS with TreeTagger

What do these clusters look like?

Table 1: Czech nominal clusters optimized towards English (`Larger`)

NOUNS CS-EN				
Cluster 0	Cluster 1	Cluster 13	Cluster 16	Cluster 12
		Fem+Sing+Nominative	Masc+Sing+Nominative	Neut+Plur+Nominative
	Fem+Sing+Vocative		Masc+Sing+Vocative	
		Fem+Sing+Accusative	Masc+Sing+Accusative	Neut+Plur+Accusative
		Fem+Sing+Genitive	Masc+Sing+Genitive	Neut+Plur+Genitive
		Fem+Sing+Dative	Masc+Sing+Dative	Neut+Plur+Dative
		Fem+Sing+Prepos	Masc+Sing+Prepos	Neut+Plur+Prepos
Fem+Dual+Instru		Fem+Sing+Instru	Masc+Sing+Instru	Neut+Plur+Instru

Table 2: Some personal pronoun clusters (`larger`)

PERSONAL PRONOUNS CS-EN	
Cluster 7	Cluster 32
Sing+Pers1+Nomin	Sing+Pers1+Accus
	Sing+Pers1+Dative
	Sing+Pers1+Prepos
	Sing+Pers1+Genitive
	Sing+Pers1+Instru

From Normalized Czech to English

Table 3: Czech-English Systems (newstest2016)

	Small System		Larger System		Largest System	
System	BLEU	OOV	BLEU	OOV	BLEU	OOV
cs2en (ali cs)	21.26	2189	23.85	1878	24.99	1246
cx2en (ali cx)	22.62 (+1.36)	1888	24.57 (+0.72)	1610	24.65 (-0.43)	988
cs2en (ali cx)	22.19 (+0.93)	2152	24.14 (+0.29)	1832	25.35 (+0.36)	1212
cx2en (ali cs)	22.34 (+1.08)	1914	24.36 (+0.51)	1627		
cx2en (100 freq)	22.82 (+1.56)	1893	24.85 (+1.00)	1614		
cx2en (lemma M sum)	22.39 (+1.13)	1860				
cx2en ($m = -10^{-4}$)			24.44 (+0.59)	1604		
cx2en ($m = 10^{-4}$)			24.05 (+0.20)	1761		
cx2en (manual)			24.46 (+0.61)	1623		

- cs2en: Moses is trained with fully inflected Czech
- cx2en: Moses with normalized Czech
- ali cs: Alignments trained with fully inflected Czech
- ali cx: Alignments trained with normalized Czech
- 100 freq: keep initial word forms for 100 most frequent words
- manual: Manual normalization (introduced earlier)

From Normalized Russian to English

Table 4: Russian-English systems (Newstest 2016)

System	BLEU	OOV
ru-en (ali ru)	19.76	2260
rx-en (ali rx)	21.02 (+1.26)	2033
rx-en (ali ru)	20.92 (+1.16)	2033
ru-en (ali rx)	20.53 (+0.77)	2048
rx-en (100 freq)	20.89 (+1.13)	2026

From Normalized Czech to French

- We now have two MRL involved.

Table 5: Czech-French systems (Newstest 2013)

System	BLEU	OOV
cs2fr (ali cs)	19.57	1845
cx2fr (ali cx)	20.19 (+0.62)	1592

2-step Translation into Czech with Morphological Reinflection

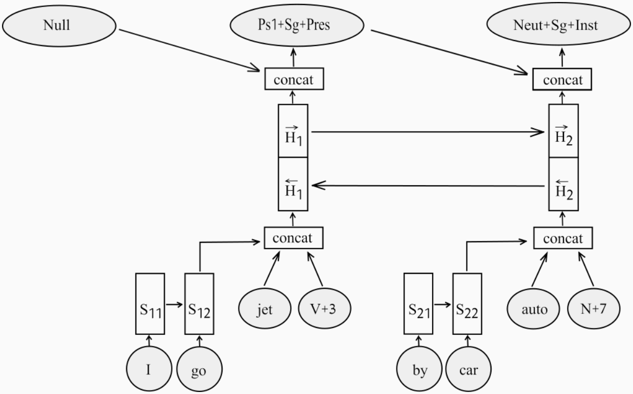

Figure 1: RNN architecture for target-side morphology prediction.

- Given a lemma and a PoS, get the word form (dictionary).

Reinflection Results

Table 6: BLEU scores for English-Czech (Newstest 2016)

	Small System			Larger System			Largest System		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
en2cs (ali cs)	15.21	0.512	0.624	17.42	0.531	0.602	19.14	0.543	0.582
en2cs (ali cx)	15.54	0.516	0.617	17.55	0.532	0.597	19.23	0.544	0.578
en2cx (1-best)	16.07	0.520	0.606	18.00	0.535	0.589	19.19	0.545	0.573
en2cx (n-best)	16.37	0.521	0.601	17.41	0.529	0.591	19.48	0.547	0.570
en2cx (nk-best)	16.93	0.525	0.602	18.81	0.540	0.588	19.95	0.548	0.572

- 1-best: 1-best MT hypothesis is reinflected
- n-best: 300-best MT hypothesis reinflected, then rescored using a LM trained with fully inflected Czech
- nk-best: same as above, add 5-best hypothesis from reinflection system.

Conclusion

Conclusion

- Providing more symmetry between analytical and synthetical languages helps to improve machine translation.
- Plain cluster IDs can be used separately and represent the grammatical content of a source word that is relevant to a target word.
- The implementation of the normalization system is available at `https://github.com/franckbrl/bilingual_morph_normalizer`.

Already performed future work: LIMSI WMT submissions

Table 7: LIMSI en2lv systems at WMT'2017

	newsdev2017	newstest2017
baseline	22.48	15.22
factored	24.19	16.36

Table 8: LIMSI en2cs systems at WMT'2017

	newstest2016	newstest2017
baseline	24.24	19.89
factored	24.59	20.54

- nmtpy system enables the prediction of two factors.
- Our system predicts normalized words (BPEs) and PoS.
- Cluster IDs are split from the lemmas (koč@@ ka+N+7 → koč@@ ka@@ N+7).
- n-best hypothesis from the factored MT system and k-best hypothesis from the dictionary are then rescored using a words-to-words system.

Thank you for your attention!

Franck Burlot, François Yvon
EAMT, Prague, Czech Republic