Two-Step MT: Predicting Target Morphology

Franck Burlot, Elena Knyazeva, Thomas Lavergne, François Yvon

LIMSI-CNRS

9 December 2016

Content

1. Introduction
2. Morphological Re-inflection
3. Impact of Data Size
4. Taking Advantage of Larger Data
5. Conclusions

Target morphology difficulties

- Dissymmetry of both languages is hard to handle:

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

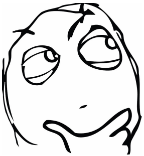
- One English word can translate into several Czech words:

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$\emptyset$).

Previous unsuccessful attempts



- Weller et al. 2013: English to French
- Weller et al. 2015: English to German
- Marie et al. 2015: same idea as Fraser 2012 (Russian)
- Allauzen et al. 2015: directly predict word forms from MT output with hidden CRF model (Russian and Romanian)

Content

1 Introduction

2 Morphological Re-inflection

3 Impact of Data Size

4 Taking Advantage of Larger Data

5 Conclusions

Re-inflection

- Normalize target side of the data:

Cs		Tags		Cx
pojedu autem	➡	pojedu auto+Neut+Sing+Inst	➡	pojedu auto+Neut+Sing+–

Re-inflection

- Normalize target side of the data:

Cs		Tags		Cx
pojedu autem	➡	pojedu auto+Neut+Sing+Inst	➡	pojedu auto+Neut+Sing+–

- Translate from English to normalized Czech:

En		Cx
I'll go by car	➡	pojedu auto+Neut+Sing+–

Re-inflection

- Normalize target side of the data:

Cs		Tags		Cx
pojedu autem	➡	pojedu auto+Neut+Sing+Inst	➡	pojedu auto+Neut+Sing+–

- Translate from English to normalized Czech:

En		Cx
I'll go by car	➡	pojedu auto+Neut+Sing+–

- Predict previously dropped tags:

En		Cx		Pred
I'll go by car	➡	pojedu auto+Neut+Sing+–	➡	pojedu auto+Neut+Sing+Inst

Re-inflection

- Normalize target side of the data:

Cs		Tags		Cx
pojedu autem	➡	pojedu auto+Neut+Sing+Inst	➡	pojedu auto+Neut+Sing+–

- Translate from English to normalized Czech:

En		Cx
I'll go by car	➡	pojedu auto+Neut+Sing+–

- Predict previously dropped tags:

En		Cx		Pred
I'll go by car	➡	pojedu auto+Neut+Sing+–	➡	pojedu auto+Neut+Sing+Inst

- Generate the word form:

Cx		Pred		Cs
pojedu auto+Neut+Sing+–	➡	pojedu auto+Neut+Sing+Inst	➡	pojedu autem

Normalization of Czech

- **Nouns**: lemma, PoS, gender and number.
- **Adjectives**: lemma, PoS, negation, degree of comparison.
- **Numerals**: *lemma, PoS*.
- **Pronouns**: lemma, PoS, subPoS, person, gender, number, number[psor], gender[psor].
- **Prepositions**: word form, PoS, case
- **Verb**: lemma and whole tag sequence
- **Adverb, interjection, conjunction, particle**: Word forms

Output re-inflection

- **Language model**: Generate all word forms and let the language model choose the most likely one using *disambig* tool (Stolcke 2002).
- **CRF**: Stacked CRF models successively predicting gender, number and case, then running a joint model using Wapiti (Lavergne 2010).
- **Greedy sequence labeller**: SVM multi-class classifier performing a greedy search (Daumé 2009).

For both latter supervised models, we also need:

- Word form generation (given a lemma and a tag sequence)
- Final disambiguation: solve remaining (mainly stylistic) ambiguities using a unigram model.

Content

1. Introduction
2. Morphological Re-inflection
3. Impact of Data Size
4. Taking Advantage of Larger Data
5. Conclusions

Experimental setup

- Ncode and Moses (contrast), 4-gram KenLMs, Mira optimization
- IWSLT'16 data (this includes WMT'16)
    - Development set: TED test 2010 + 2011
    - Test set: Ted test 2012 + 2013
    - Parallel data:
        - First 10k from TED training set
        - Full TED set (117k)
        - + QED (242k)
        - + europarl (885k)
        - + news-commentary (1M)
    - Monolingual data (various subsets ranging from 5M to 200M):
        - Target side of the biggest parallel corpus
        - Czeng-1.6-pre subtitles
        - news corpora (WMT'16)
        - common-crawl (WMT'16, filtered)

Growing parallel data

Data	Moses			
	en2cs	LM	CRF	Greedy
10k	10.06	9.96 (-0.10)	11.60 (+1.54)	**11.64** (+1.58)
117k	15.70	15.20 (-0.50)	16.70 (+1.00)	**16.78** (+1.08)
242k	15.96	15.32 (-0.64)	16.72 (+0.76)	**16.90** (+0.94)
885k	16.75	16.45 (-0.30)	17.74 (+0.99)	**17.94** (+1.19)
1M	17.14	16.51 (-0.63)	17.64 (+0.50)	**17.88** (+0.74)

Data	Ncode			
	en2cs	LM	CRF	Greedy
10k	10.62	10.44 (-0.18)	12.13 (+1.51)	**12.28** (+1.56)
117k	15.77	15.52 (-0.25)	17.17 (+1.40)	**17.32** (+1.55)
242k	16.06	15.68 (-0.38)	17.17 (+1.11)	**17.32** (+1.26)
885k	16.94	16.67 (-0.27)	18.04 (+1.10)	**18.25** (+1.29)
1M	17.15	16.64 (-0.51)	17.99 (+0.84)	**18.13** (+0.98)

BLEU scores over en2cs and en2cx2cs

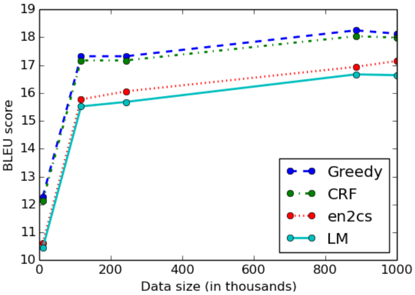
Growing monolingual data

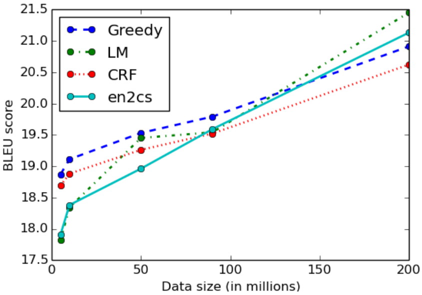
Data	Moses			
	en2cs	LM	CRF	Greedy
5M	18.01	18.05 (+0.04)	18.73 (+0.72)	**18.84** (+0.83)
10M	18.58	18.42 (-0.16)	18.87 (+0.29)	**19.05** (+0.47)
50M	18.97	19.19 (+0.22)	19.02 (+0.05)	**19.22** (+0.25)
90M	19.34	19.40 (+0.06)	19.26 (-0.08)	**19.51** (+0.17)
200M	20.71	**20.81** (+0.10)	19.75 (-0.96)	20.02 (-0.69)

Data	Ncode			
	en2cs	LM	CRF	Greedy
5M	17.91	17.82 (-0.09)	18.69 (+0.78)	**18.87** (+0.96)
10M	18.38	18.34 (-0.04)	18.88 (+0.50)	**19.11** (+0.72)
50M	18.96	19.45 (+0.49)	19.26 (+0.30)	**19.53** (+0.57)
90M	19.59	19.54 (+0.05)	19.52 (-0.07)	**19.79** (+0.20)
200M	21.13	**21.45** (+0.32)	20.62 (-0.51)	20.91 (-0.22)

BLEU scores over en2cs and en2cx2cs

Model Comparison





Scores for re-inflection using different models over increasing parallel and monolingual data size.

Content

1. Introduction
2. Morphological Re-inflection
3. Impact of Data Size
4. Taking Advantage of Larger Data
5. Conclusions

N-best hypothesis re-inflection

- Re-inflection with 1-best hypothesis: fixed set of words, fixed order
- Re-inflection can take advantage of the diversity provided by n-best hypothesis

N-best hypothesis are re-inflected and given a new score with an LM trained on fully inflected Czech. All scores (Translation step and LM) are interpolated using Mira. Two kinds of LM:

- N-gram LM (KenLM)
- Neural LM with characted-based word representation

We also take the k-best CRF predictions, leading to nk-best hypothesis.

N-best hypothesis re-inflection

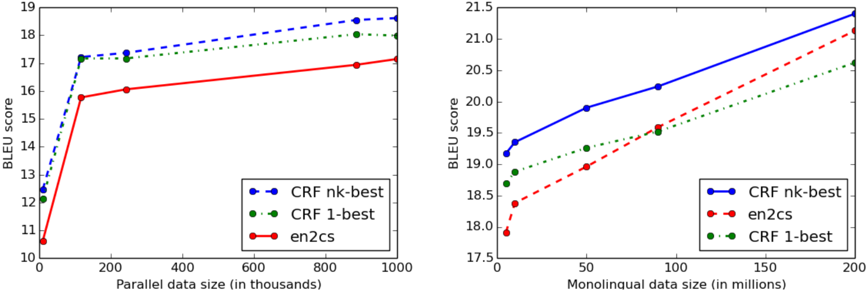
Model	**10k/10k**	**117k/117k**	**242k/242k**	**885k/885k**	**1M/1M**
en2cs	10.62	15.77	16.06	16.94	17.15
LM	10.42 (-0.20)	15.47 (-0.30)	15.81 (-0.25)	16.64 (-0.30)	16.72 (-0.43)
CRF	12.39 (+1.77)	17.31 (+1.54)	17.17 (+1.11)	18.24 (+1.30)	18.23 (+1.08)
+ CRF k-best	**12.47** (+1.85)	17.22 (+1.45)	17.37 (+1.31)	**18.55** (+1.61)	**18.62** (+1.47)
Greedy	12.39 (+1.77)	**17.49** (+1.72)	**17.65** (+1.59)	18.31 (+1.37)	18.55 (+1.40)
Model	**885k/5M**	**885k/10M**	**885k/50M**	**885k/90M**	**885k/200M**
en2cs	17.91	18.38	18.96	19.59	21.13
LM	17.91 (+0.00)	18.30 (-0.08)	19.20 (+0.24)	19.81 (+0.22)	21.29 (+0.16)
CRF	18.81 (+0.90)	19.23 (+0.85)	19.50 (+0.54)	20.02 (+0.43)	21.07 (-0.06)
+ CRF k-best	19.17 (+1.26)	19.35 (+0.97)	**19.90** (+0.94)	**20.24** (+0.65)	**21.40** (+0.27)
Greedy	**19.23** (+1.32)	**19.54** (+1.16)	19.84 (+0.88)	20.23 (+0.64)	21.35 (+0.22)

BLEU scores over en2cs and en2cx2cs (Ncode)

Setup	TED-2015	TED-2016	QED-2016
en2cs baseline	18.37	15.27	16.20
N-gram LM	19.65 (+1.28)	16.63 (+1.36)	16.25 (+0.05)
WE	19.65 (+1.30)	16.66 (+1.39)	**16.26 (+0.06)**
CWE	**19.77 (+1.42)**	**16.80 (+1.53)**	15.96 (-0.24)
CWE-CWE	19.25 (+0.88)	16.31 (+1.04)	15.27 (-0.93)

BLEU scores for re-ranked re-inflected nk-best translation hypothesis (en2cx2cs) over the official IWSLT 2016 test sets

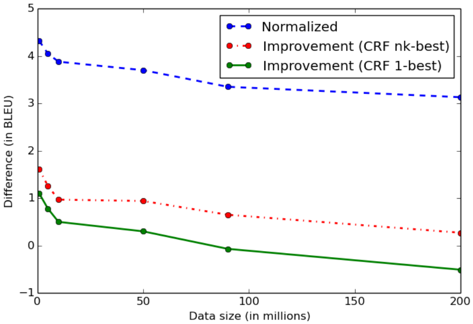
Ranking Comparison



Scores for CRF re-inflexion of 1-best and nk-best hypothesis over increasing parallel and monolingual data size.

Source	I will bypass you
CRF 1-best	budu **ti** obejít
	will you-Dative bypass-Perfective
CRF nk-best	budu **tě** obcházet
	will you-Accusative bypass-Imperfective

Ranking Comparison



Difference in BLEU score between baseline (cs) and both normalized (cx) and re-inflected outputs (cx2cs) with growing monolingual data.

Content

1 Introduction

2 Morphological Re-inflection

3 Impact of Data Size

4 Taking Advantage of Larger Data

5 Conclusions

What do these experiments show?

- Re-inflection is more effective in low-ressource conditions
- Less, but still effective when vast amounts of monolingual data available (LM re-inflection and / or n-best re-scoring)
- 885k/200M system generates 6.82% of word types not seen in training data (1.76% tokens)

There is a right model for each data setup:

- Weller et al. 2013 got no improvement with CRF re-inflection on en2fr (9M/32M)
- Same for Marie et al. 2015 on en2ru (2.3M/46M)
- Fraser 2012 got no improvement with n-best re-inflection on en2de (1.5M/10M)

Future work:

- Manual normalization is not optimal, how can this be done automatically?
- Strategies to lower dependency on human informed ressources quality (tagger, dictionary)
- How does re-inflection perform with neural MT?

Thank you for your attention!