Helsinki, Turku and Uppsala @ WMT

Jörg Tiedemann & Robert Östling, University of Helsinki

Fabienne Cap & Sara Stymne, Uppsala University

Filip Ginter & Jenna Kanerva, University of Turku

Factored Phrase-Based Models

Multiple translation paths + morphological generation

- translation of surface words
- open-class words replaced with lemmas
- translation of morphological features

Factored Phrase-Based Models

Multiple translation paths + morphological generation

- translation of surface words
- open-class words replaced with lemmas
- translation of morphological features

Re-inflection models

- Sara will tell you more ...

Factored Phrase-Based Models

Multiple translation paths + morphological generation

- translation of surface words
- open-class words replaced with lemmas
- translation of morphological features

Re-inflection models

- Sara will tell you more ...

More data

- OPUS, back-translations, crawled data

More Data and Efficient Alignment

newstest 2015	BLEU	time for wordalign real	CPU
GIZA++	13.65	38,514s	–
fast_align	13.56	682s	8,344s
efmaral	14.10	370s	895s
+ OPUS	14.81	–	–
+ alternatives	15.55	2,630s	6,599s
+ WWW-LM	16.98	–	–
retuned	18.11	–	–
back-translated	14.78	954s	2,606s
+ OPUS, ...	18.22	2,758s	7,187s

More Data and Efficient Alignment

newstest 2015	BLEU	time for wordalign real	time for wordalign CPU
GIZA++	13.65	38,514s	–
fast_align	13.56	682s	8,344s
efmaral	14.10	370s	895s
+ OPUS	14.81	–	–
+ alternatives	15.55	2,630s	6,599s
+ WWW-LM	16.98	–	–
retuned	18.11	–	–
back-translated	14.78	954s	2,606s
+ OPUS, ...	18.22	2,758s	7,187s

More Data and Efficient Alignment

newstest 2015	BLEU	time for wordalign real	time for wordalign CPU
GIZA++	13.65	38,514s	–
fast_align	13.56	682s	8,344s
efmaral	14.10	370s	895s
+ OPUS	14.81	–	–
+ alternatives	15.55	2,630s	6,599s
+ WWW-LM	16.98	–	–
retuned	18.11	–	–
back-translated	14.78	954s	2,606s
+ OPUS, ...	18.22	2,758s	7,187s

More Data and Efficient Alignment

newstest 2015	BLEU	time for wordalign real	time for wordalign CPU
GIZA++	13.65	38,514s	–
fast_align	13.56	682s	8,344s
efmaral	14.10	370s	895s
+ OPUS	14.81	–	–
+ alternatives	15.55	2,630s	6,599s
+ WWW-LM	16.98	–	–
retuned	18.11	–	–
back-translated	14.78	954s	2,606s
+ OPUS, ...	18.22	2,758s	7,187s

New Types of Data Sets

Synthetic training data

- Translate monolingual Finnish to English (using SMT)
- Morphology, compound splitting, placeholder prepos.
- Extremely useful in Neural MT, works also in PB-SMT

New Types of Data Sets

Synthetic training data

- Translate monolingual Finnish to English (using SMT)
- Morphology, compound splitting, placeholder prepos.
- Extremely useful in Neural MT, works also in PB-SMT

Alternative movie subtitles

Mitä on tekeillä ?
- Mitä ihmettä te teette ?

Ei tässä hätää .
En tiennyt , olen pahoillani .

Mutta tajuan kyllä tyttöjäkin .
Mutta ymmärrän kyllä tyttöjäkin .

New Types of Data Sets

Synthetic training data

- Translate monolingual Finnish to English (using SMT)
- Morphology, compound splitting, placeholder prepos.
- Extremely useful in Neural MT, works also in PB-SMT

Alternative movie subtitles

Mitä on tekeillä ?
- Mitä ihmettä te teette ?

Ei tässä hätää .
En tiennyt , olen pahoillani .

Mutta tajuan kyllä tyttöjäkin .
Mutta ymmärrän kyllä tyttöjäkin .

Test set with two reference translations

Gappy Language Models

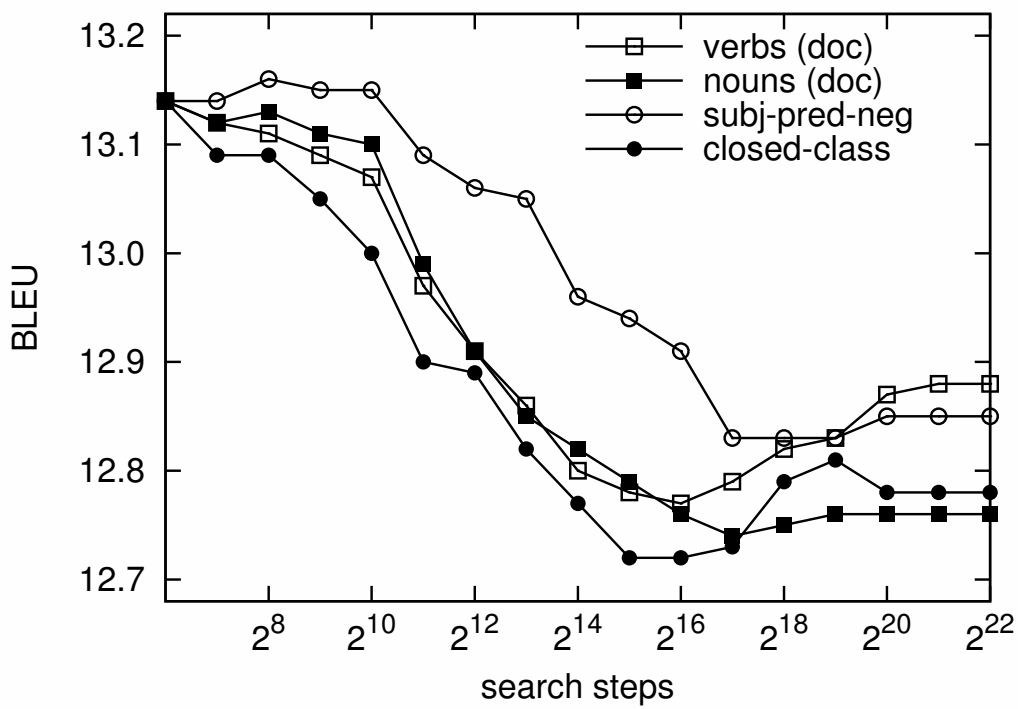
Document-level decoding with Docent

- n-gram language models over selected words
- selection based on PoS and dependency relations
- agreement issues even across sentence boundaries

Gappy Language Models

Document-level decoding with Docent

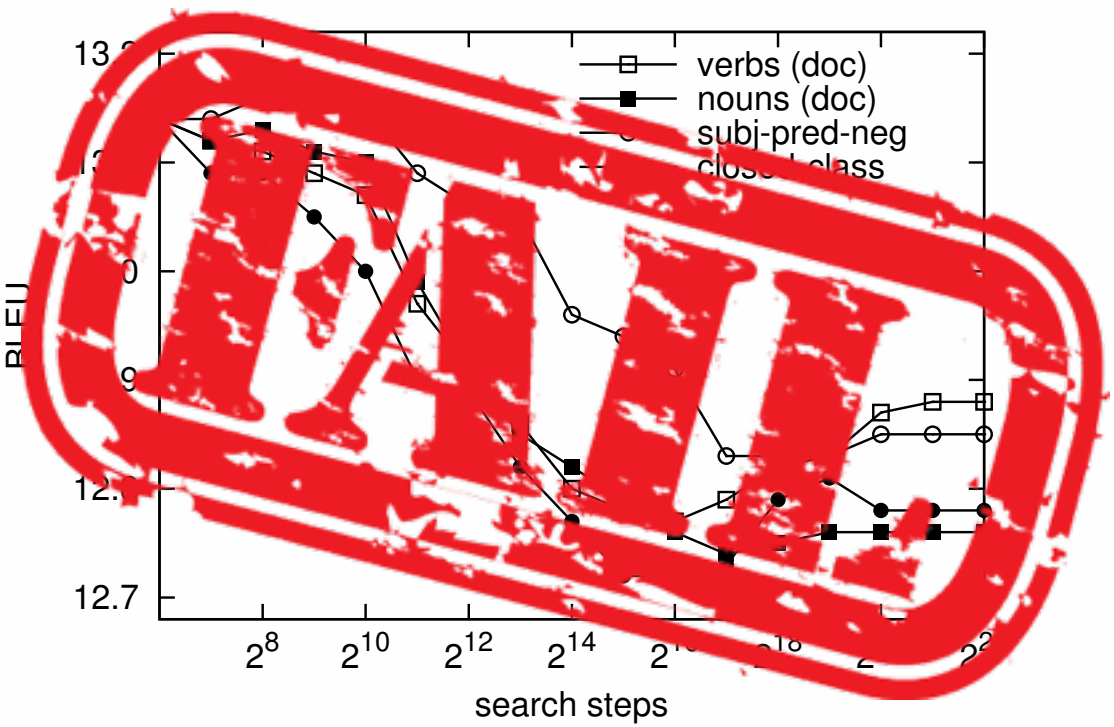
- n-gram language models over selected words
- selection based on PoS and dependency relations
- agreement issues even across sentence boundaries



Gappy Language Models

Document-level decoding with Docent

- n-gram language models over selected words
- selection based on PoS and dependency relations
- agreement issues even across sentence boundaries



FAIL

WMT 2016 Results

English – Finnish	BLEU lower	BLEU cased	TER	unknown words #tokens	unknown words #types
constrained - basic	13.3	12.7	0.782	1,582	862
constrained - factored	13.5	12.8	0.784	1,659	1,233
constrained - basic + back-translated	14.2	13.6	0.770	**1,024**	**649**
constrained + factored + back-translated	**14.3**	**13.6**	**0.765**	1,103	890
constrained - re-inflection	12.2	11.6	0.793		
unconstrained - basic	17.0	16.2	0.746	**124**	**60**
unconstrained - factored	16.6	15.7	**0.744**	804	593
unconstrained - basic + back-translated	**17.1**	**16.4**	0.752	544	305

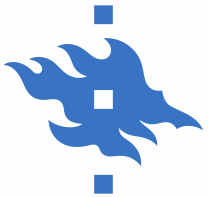
Finnish – English	BLEU lower	BLEU cased	TER	unknown words #tokens	unknown words #types
constrained - factored	20.5	19.3	0.706	2,655	2,004
unconstrained - factored	**23.3**	**22.1**	**0.670**	**1,128**	**842**

Example Translations

Input: A 9-year-old boy missing at Tampere was found
SMT: 9-vuosi Tampereella kadonnut vanhus löytyi
Reference: *Tampereella karkuteillä ollut 9-vuotias poika löytyi*

Input: The police organised a search at Tampere on Tuesday evening because a 9-year old boy ran away.
SMT: Poliisi järjesti etsinnän Tampereella tiistai-iltana, koska 9-vuotias poika juoksi karkuun.
Reference: *Poliisilla on ollut etsintätehtävä 9-vuotiaan pojan karkumatkan takia Tampereella tiistaina illalla.*

Input: The boy started off from western Tampere with a friend, and a missing person notice spread in social media.
SMT: Poika aloitti Länsi-Tampereella pois ystävän kanssa, ja kadonneen henkilön ilmoitus levisi sosiaalisessa mediassa.
Reference: *Länsi-Tampereelta kaverinsa kanssa liikkeelle lähtenyttä poikaa koskeva katoamisilmoitus on levinnyt sosiaalisessa mediassa.*



Comparison to Neural MT

Input: A 9-year-old boy missing at Tampere was found
Neural MT: Pikkupoika löytyi Tampereelta
Reference: *Tampereella karkuteillä ollut 9-vuotias poika löytyi*

Input: The police organised a search at Tampere on Tuesday evening because a 9-year old boy ran away.
Neural MT: Poliisi järjesti etsinnät Tampereella tiistaiiltana, koska vanha poika juoksi karkuun.
Reference: *Poliisilla on ollut etsintätehtävä 9-vuotiaan pojan karkumatkan takia Tampereella tiistaina illalla.*

Input: The boy started off from western Tampere with a friend, and a missing person notice spread in social media.
Neural MT: Poika lähti liikkeelle Tampereen länsilaidalla ystävänsä kanssa, ja kadonneen henkilön levisi sosiaalisessa mediassa.
Reference: *Länsi-Tampereelta kaverinsa kanssa liikkeelle lähtenyttä poikaa koskeva katoamisilmoitus on levinnyt sosiaalisessa mediassa*