Morphological Segmentation for Machine Translation

in Elx, 2014

Tommi A Pirinen ‹tommi.pirinen@computing.dcu.ie›

DCU, CNGL
Abumatran

August 19, 2014

About myself

- BSc in CS from U Joensuu (now UEF.fi) 2004, MA in Comp Ling from U Helsinki 2008 and PhD in Comp Ling from U Helsinki 2014
- CL projects such as: open source morphology of Finnish, giellatekno comp ling repo, HFST, apertium (fin-eng,swe,hun,hbs,ces,gle,rus etc.)
- Other FLOSS projects: Gentoo Linux, Finnish localisation, probably more

Background: Morphological complexity

- debated, hot topic in linguistics
- for MT measuring is relatively simple:
- number of unique tokens / oov tokens per dataset, etc.
- e.g., most English nouns have ~4 forms (plural + possessives) Serbo-Croatian at least 14 (cases, plurals), Finnish at least 5,271 (cases, plurals, possessives, clitics,' allomorphs)
- with compounding and derivation vocabulary is practically endless
- So, amount of data for statistical models might need to be more

Motivation / Use cases

i.e., what is it good for:

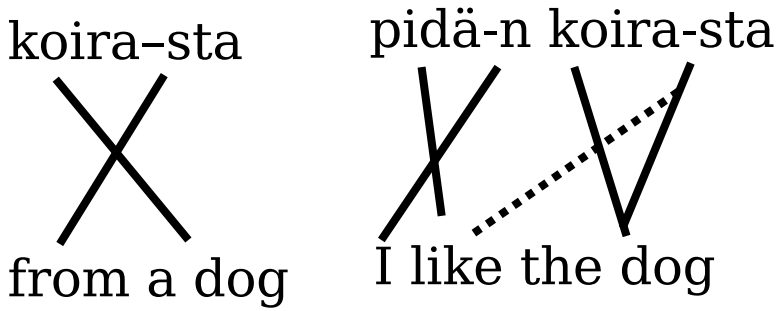
- Morphology-aware terminology management in localisation applications
- there's practically very little localisation software that handles morphology with terms, names etc. well
- e.g., software localisation I've seen so far doesn't quite allow for morphology at all (gettext etc.)
- e.g., translation memories will fuzzy match copy the most common forms forcing translators to post editorially inflect all terms

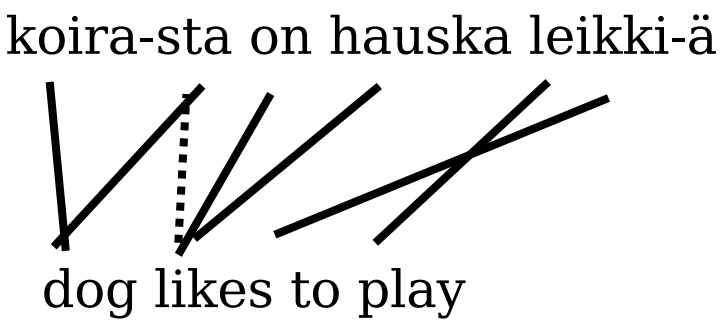
Morphological segmentation

- For MT: ideally break wordform down to morphs that would translate to words in the other language
- e.g. talo : talo/ssa : talo_i/hin : puu/talo etc. for house : in house : to house_s : wooden house (Finnish and English share plural as suffix)
- segmentation can be ambiguous, e.g. katos/ta ~ kato/sta (between katto and katos)
- with rulebased morphology we can select these rather easily
- statistical approach such as Morfessor uses minimum description length with corpora learning to segment the likely morphs to minimize number of unique morphs in whole data; may need tuning and data selection
- N.B. alternative to morphological segmentation would be to use factored models with morphological analyses and lemmatization or stemming

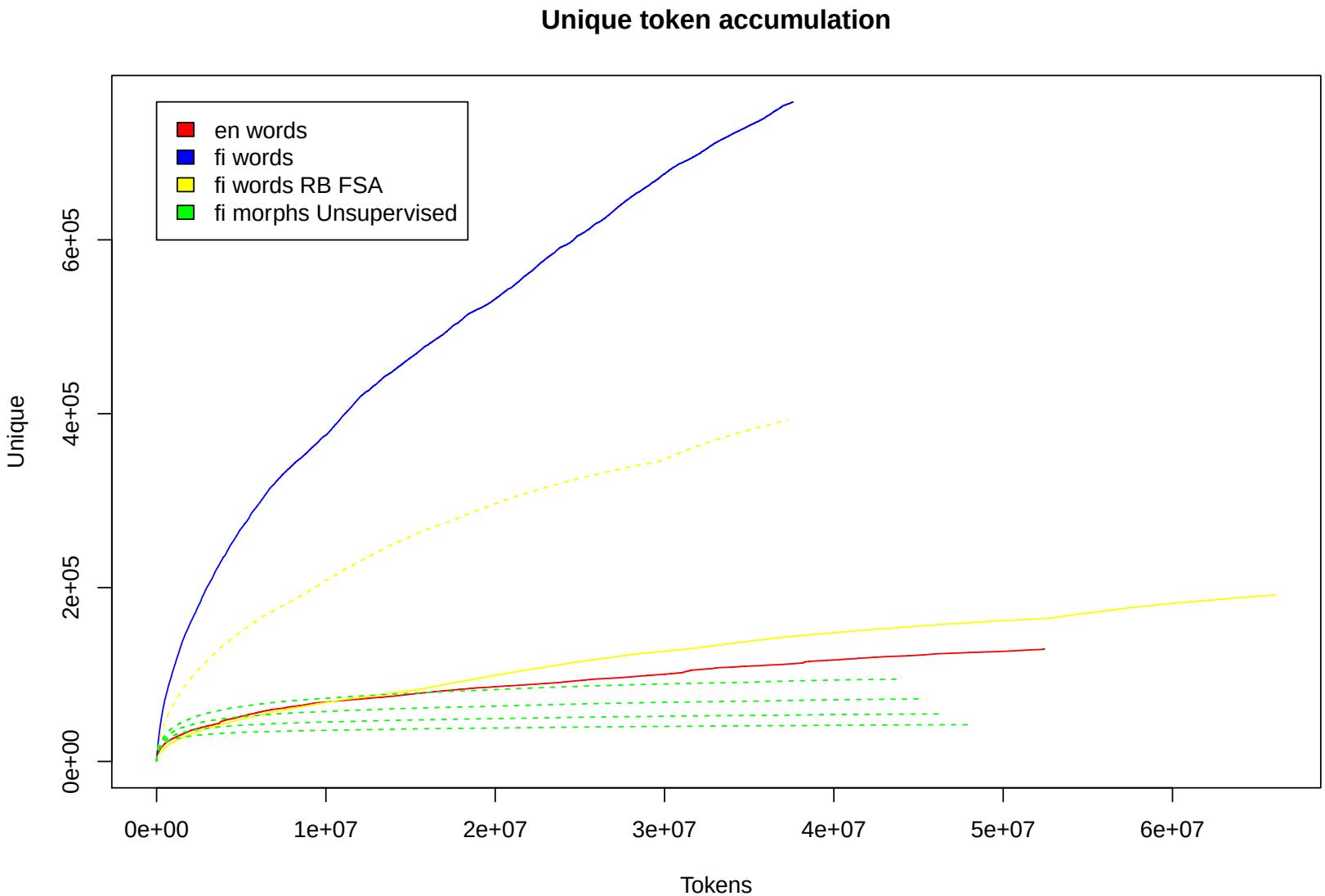
Segmentations in Translation

When semantic case suffixes are used prototypically they align 1:1 English. When semantic cases are e.g. object cases of the verb in a construction they can be more complex though.





Europarl statistics for English, Finnish and two segmentations



Implementation: Baseline Moses

This is starting point for all my experiments

- as per `http://www.statmt.org/moses/?n=Moses.Baseline`
- Split corpus, tokenize, truecase, clean
- train Language models
- align, train translation model
- tuning
- test set is tokenised, truecased, translated, and compared

Morphological segmentations

Translating *from* morphologically complex language (e.g., Fin $\rightarrow$ Eng)

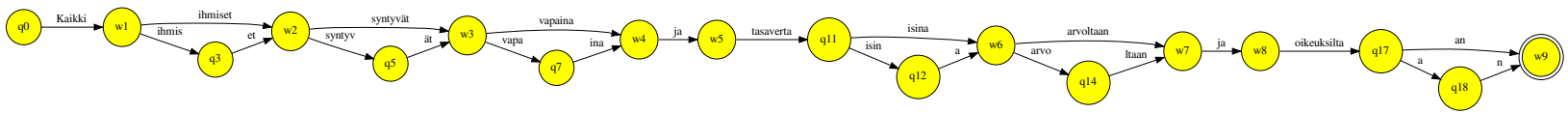
- training corpus is segmented on source language side (e.g., Fin) before training translation model
- test set needs to be segmented before translating
- morph lattices can be used instead of 1-best segmentation

Translating *into* morphologically complex language (e.g., Eng $\rightarrow$ Fin)

- training corpus is segmented on target language side (e.g., Fin) before training language and translation models
- translation need to be joined before evaluating
- what to do with stray segments and illegal combinations

Morph lattices

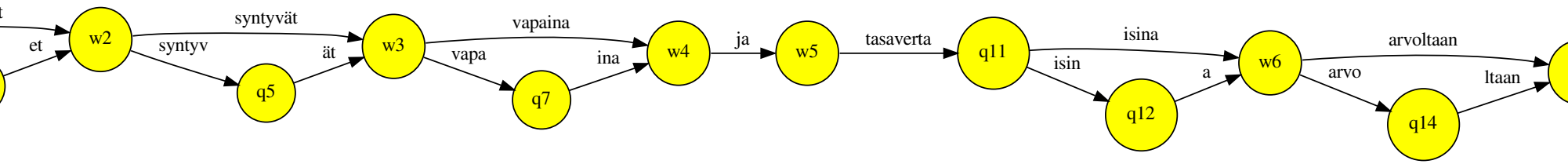
Ambiguous segmentations and statistical n-best segmentations can be fed to moses as if they were word lattices:



(from morfessor's 2-best segmentation; morphologically correct: "Kaikki ihmise/t synty/vät vapa/i/na ja tasa/vertais/i/na arvo/lta/an ja oikeuks/i/lta/an")

Morph lattices

Ambiguous segmentations and statistical n-best segmentations can be fed to moses as if they were word lattices:



Morphological segmentations and statistics

- Rule-based segmentation needs either analyser trained with good statistics or segmentation trained with good segmentations
- Some morfessor algorithms can be trained with gold data
- There's not much pre-segmented data available
- currently what I did is use automatic segmentations of different sources to give some statistics (or weights of WFSA analyser if available)

The Experiments

Writing comparison of results with different parameters:

- Different segmentation algorithms: Rule-based (morphs, words), Morfessor (2.0, categories map, ml)
- Different size of training data for segmenters: europarl partitions (full, half, quarter, eighth)
- Different size of training data for translation model: same europarl partitions
- Joining algorithms

Very Preliminary Experiments

WIP as of August 19, 2014:
English to Finnish

Score: Model	**BLEU** (WPT 05)	**TER** (WPT 05)	**BLEU** (UNDHR)	**TER** (UDHR)
Apertium Baseline	0.03		0.01	0.96
Moses Baseline	0.15		0.18	0.71
Moses Baseline 1/4	0.14		0.11	0.71
Moses 1-Best Compounds (FSA)	0.15		0.17	0.70
Moses 1-Best Morphs (FSA + Morfessor)	0.14		0.10	0.74

More Experiments Under Way

WIP as of August 19, 2014:
Finnish to English

Score: Model	**BLEU** (WPT 05)	**TER** (WPT 05)	**BLEU** (UNDHR)	**TER** (UNDHR)
Apertium Baseline	0.05		0.05	
Moses Baseline	0.22		0.28	
Moses Baseline 1/4	0.21		0.24	
Moses 1-Best Morphs (Morfessor)	0.22		0.24	

Future World

- Include more languages in tests: South Slavic, etc.
- Limit segmentation to rare words instead of all
- Error analysis should reveal some interesting things
- Segmentation gold standards
- ...(suggestions welcome)

URLs and references

- Tommi.Pirinen@computing.dcu.ie
- http://www.computing.dcu.ie/~tpirinen/
- http://www.github.com/flammie/purplemonkeydishwasher/
- SVN at redmine repo mt-development