Morphology in Statistical Machine Translation

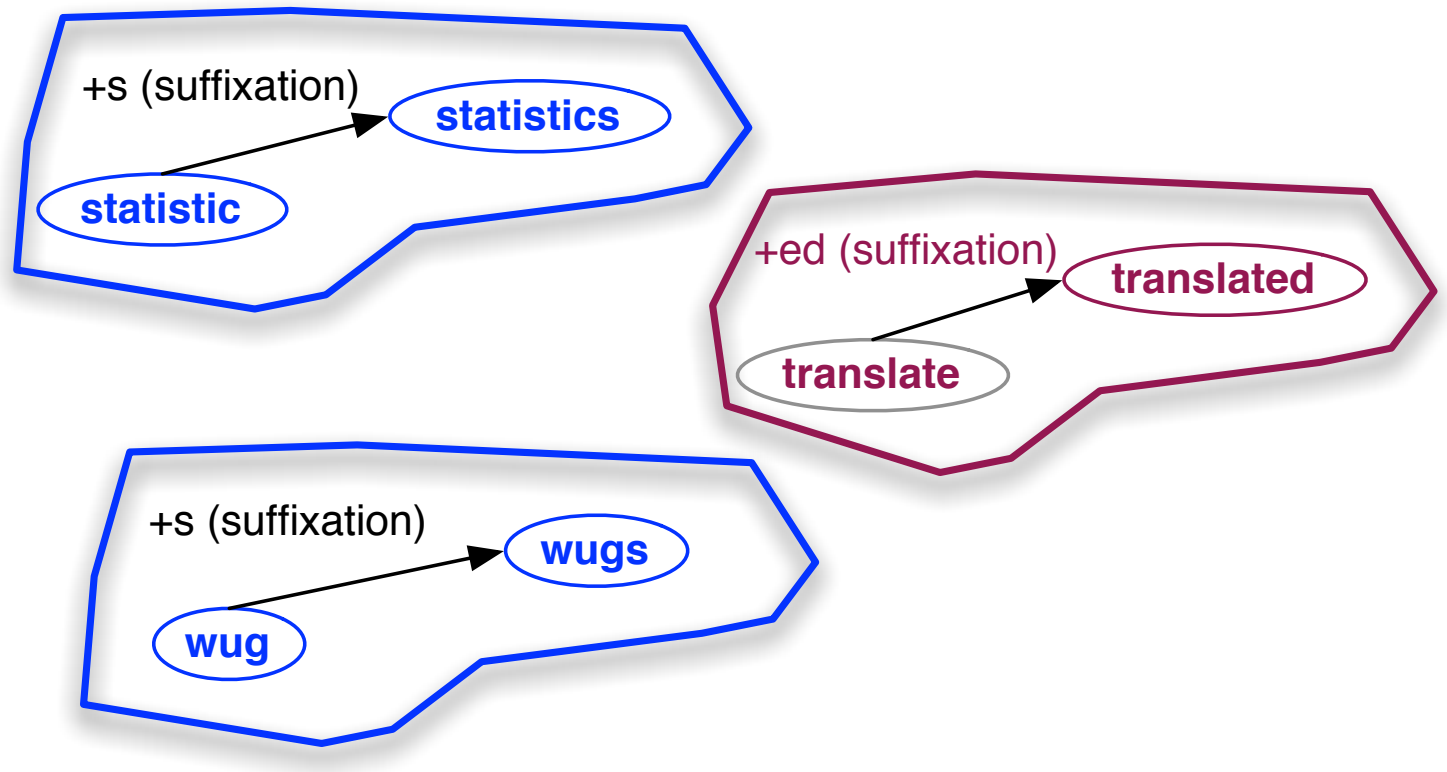
Philip Williams
University of Edinburgh

MTM 2014
12 September 2014

Outline

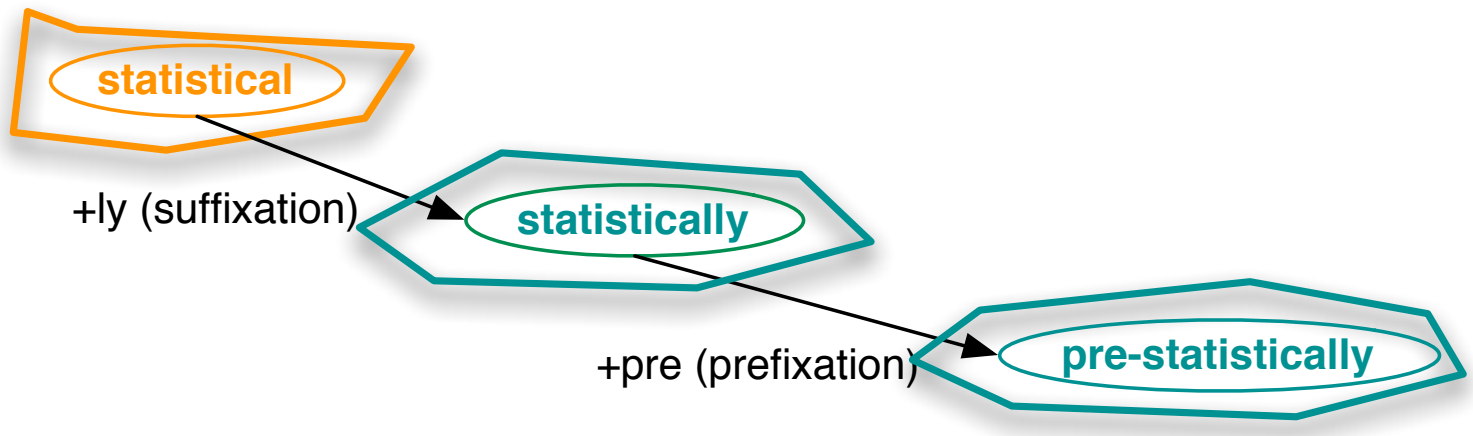
Outline

Morphology — Some Examples in English



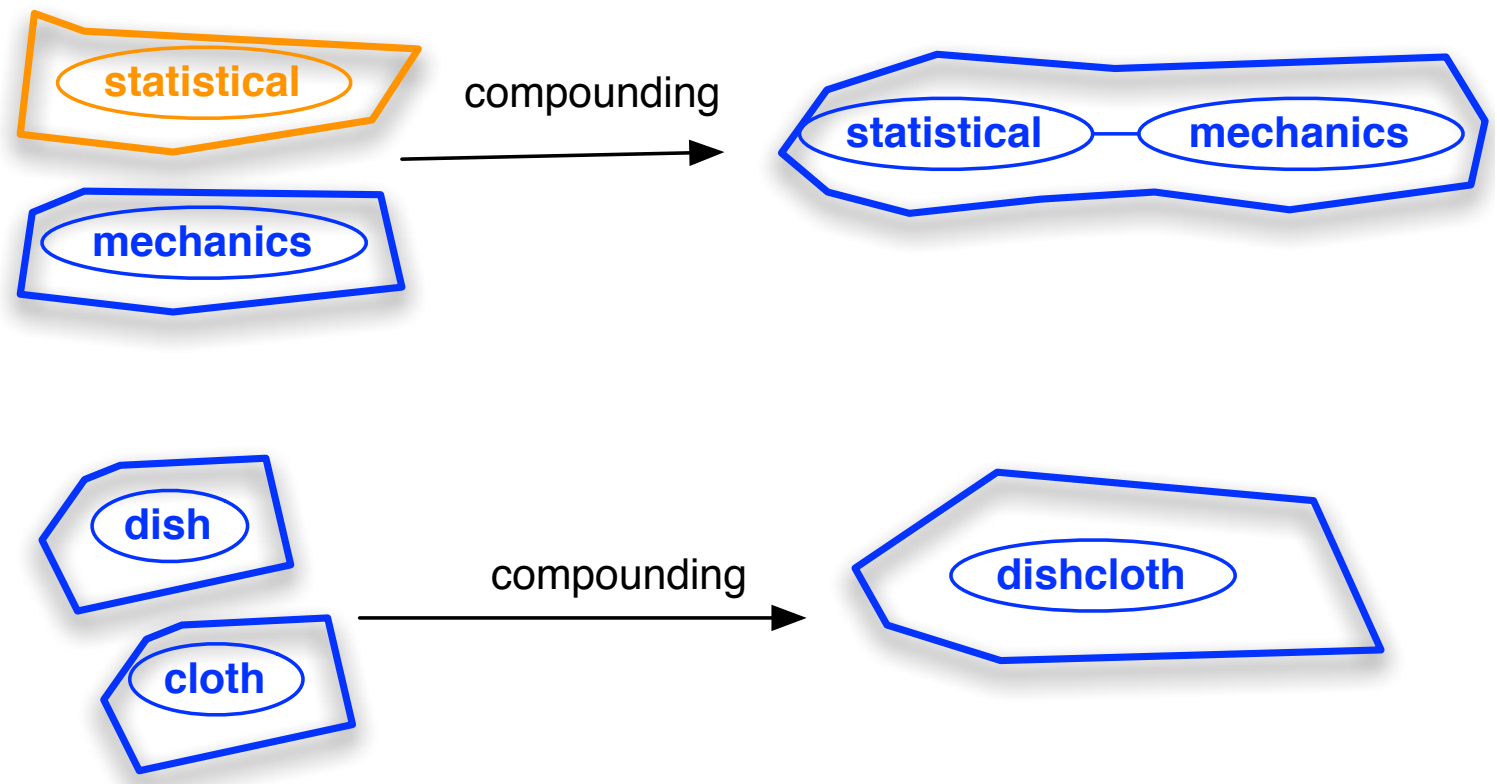
- Inflection: forms vary to fit grammatical context
  (the grammatical features here are number and tense)

Morphology — Some Examples in English



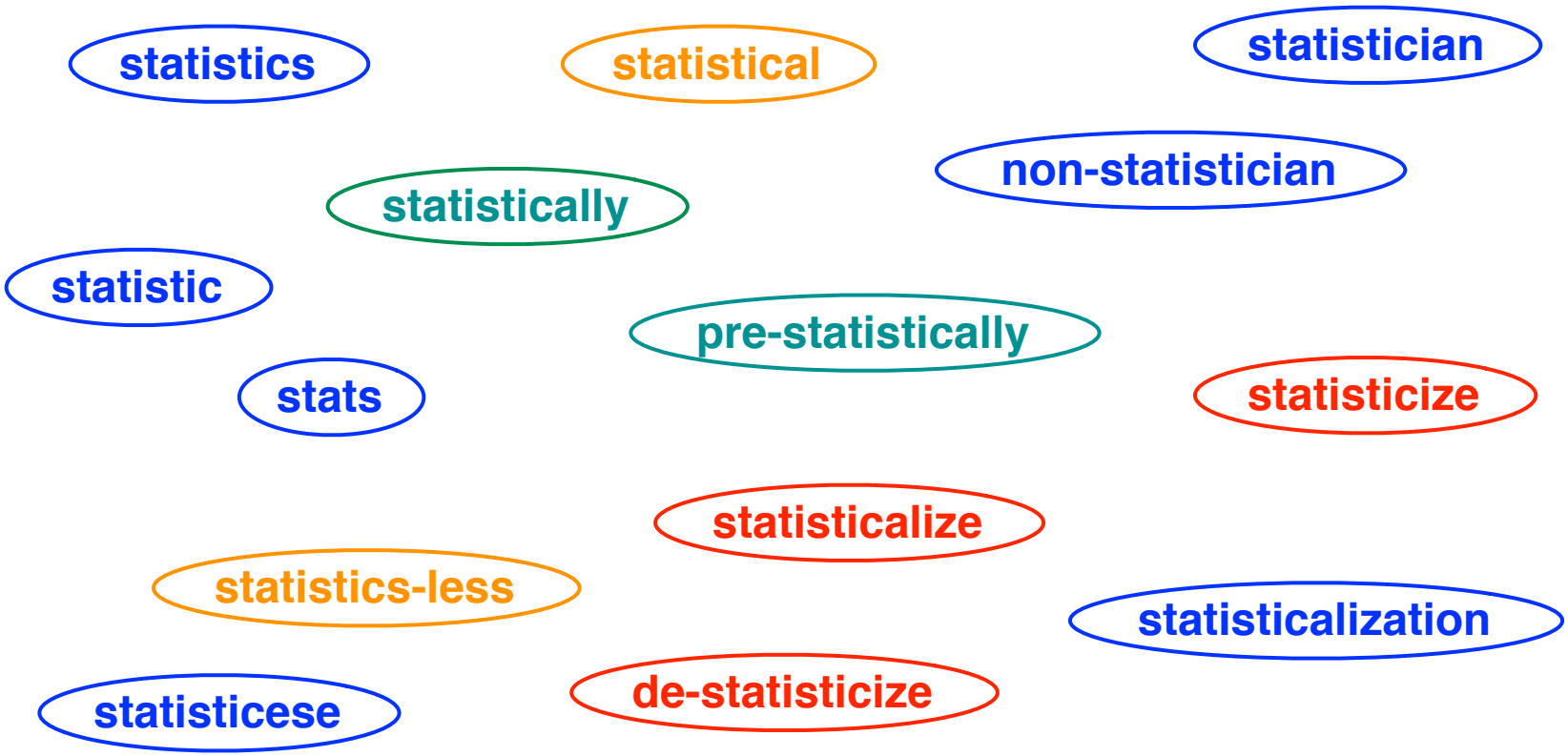
- Derivation: change in meaning and/or category (meaning change is somewhat predictable)

Morphology — Some Examples in English



- Compounding: combination of complete word forms (sometimes produces MWEs, sometimes single forms)

Morphology — The Traditional SMT Approach



Morphology in Other Languages

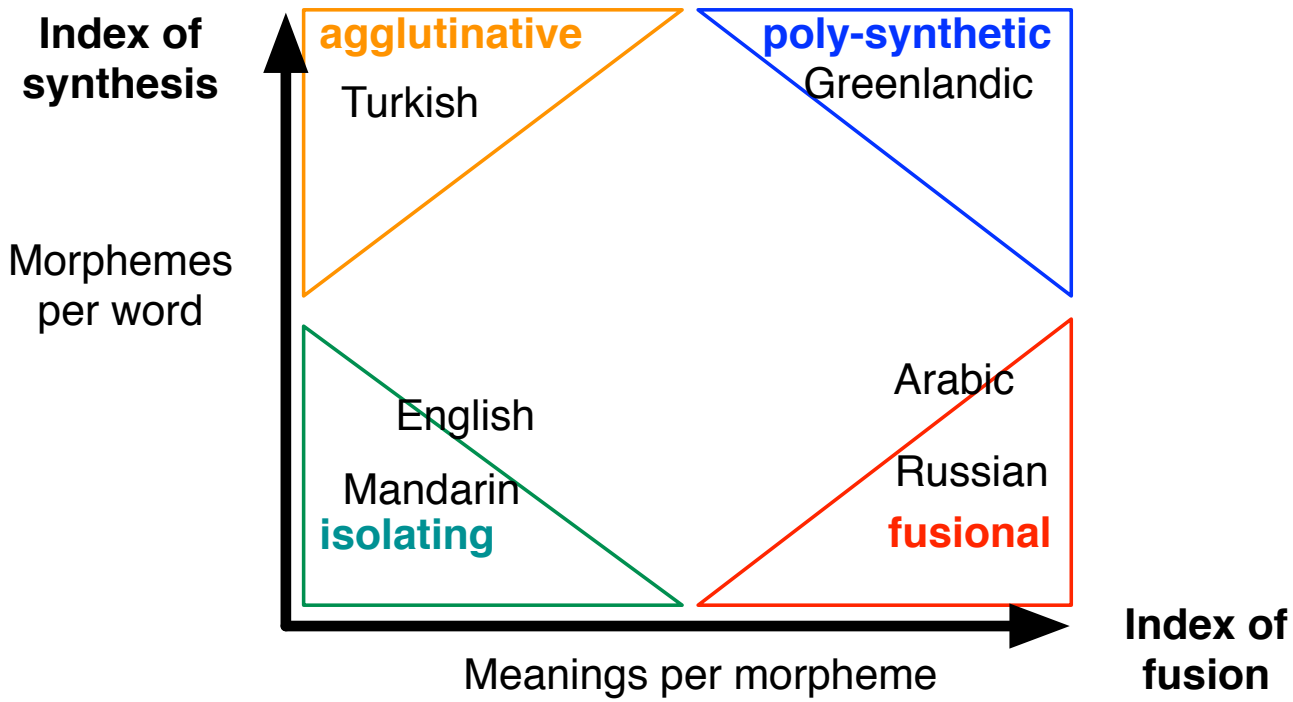
си́ний - blue

	Singular			Plural
	Masc.	Fem.	Neut.	
Nominative	си́ний	си́няя	си́нее	си́ние
Genitive	си́него	си́ней	си́него	си́них
Dative	си́нему	си́ней	си́нему	си́ним
Accusative	N or G	си́нюю	си́нее	N or G
Instrumental	си́ним	си́ней	си́ним	си́ними
Prepositional	си́нем	си́ней	си́нем	си́них

Morphology in Other Languages

Turkish	English
Avrupa	Europe
Avrupalı	of Europe / European
Avrupalılaş	become European
Avrupalılaştır	to cause to become European / Europeanize
Avrupalılaştırama	be unable to Europeanize
Avrupalılaştıramadık	we were unable to Europeanize

Morphology in Other Languages



Morpheme - "The smallest unit of morphology that has its own meaning"

Why Morphology Matters for SMT

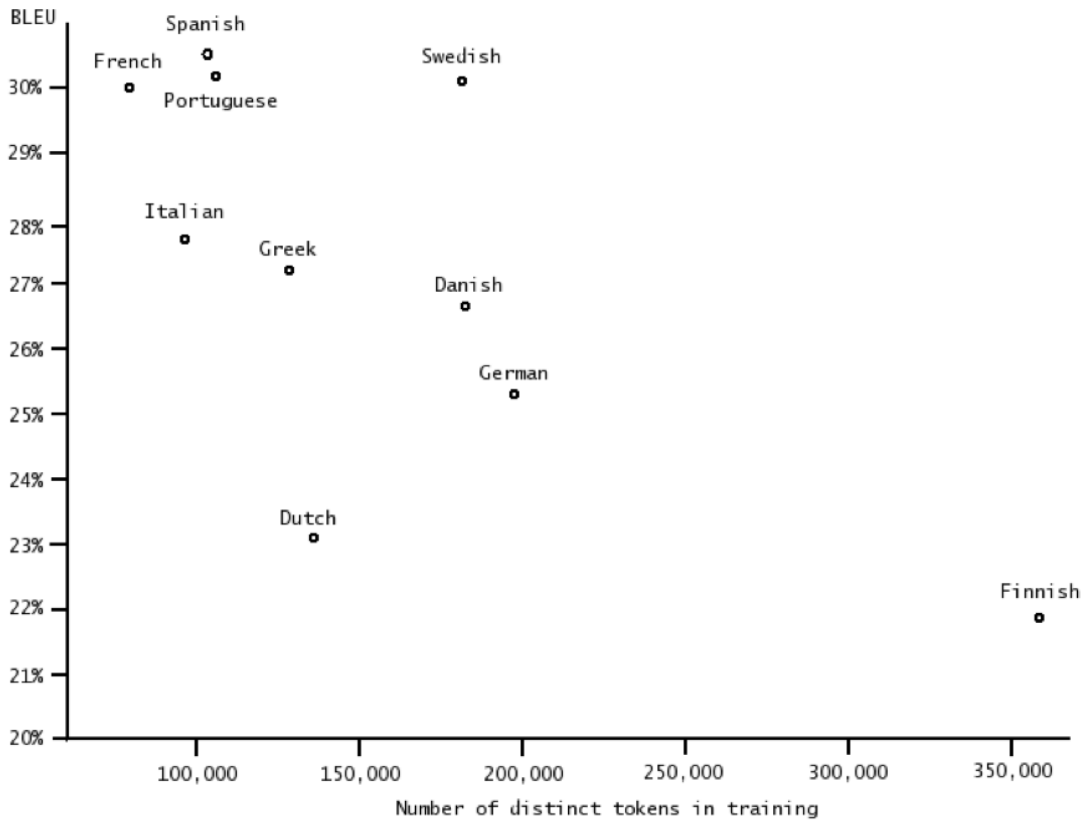


Figure 4: Vocabulary size vs. BLEU score when translating into English (which has about 65,000 distinct word forms)

(From Koehn (2005))

A Rough Classification of Subtasks

1. **Segmentation**
   - In a nutshell: split word forms to overcome sparseness
2. **Simplification**
   - In a nutshell: simplify word forms to improve source-target symmetry
3. **Feature coherence**
   - In a nutshell: produce word forms that express coherent feature values (consistent with the source, and consistent across target words)

(Some approaches perform one of these tasks; others perform a combination)

A Rough Classification of Integration Strategies

1. **Pre-processing and/or post-processing**
   - In a nutshell: pre-process training data to better fit standard SMT models and/or post-process translations
2. **Add Scoring Models**
   - In a nutshell: add morphologically-aware feature functions to score translation candidates
3. **Enrich the Translation Rules**
   - In a nutshell: use morphological information to add or remove translation candidates to or from the search space

Outline

Example 1 (source-side morphology, supervised)
German Compound Splitting (Koehn and Knight, 2003)

Problem German compounding frequently introduces new word forms leading to data sparsity issues

Example Aktionsplan = Aktion+s+Plan (action plan)
Fahrpreisermäßigung = Fahrpreis+Ermäßigung (fare reduction)

Outline

1. Learn a compound splitting model from the training data
2. Split compounds on the souce-side of the parallel corpus and in input sentences
3. Train a standard SMT model

Example 1 (source-side morphology, supervised)
German Compound Splitting (Koehn and Knight, 2003)

- Several splitting models: max split, count-based model, use of aligned English sentences to guide split
- Units must have been observed as separate words
- Compare accuracy of models on manually-annotated test set and in MT
- Highest accuracy model not best for phrase-based MT

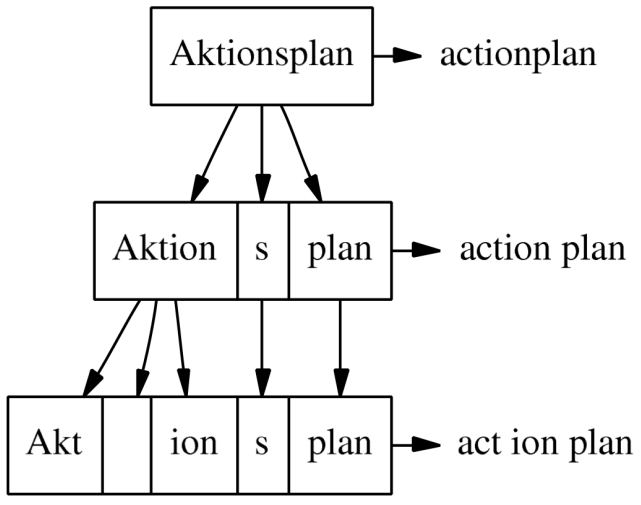


Figure 1: Splitting options for the German word *Aktionsplan*

Example 2 (source-side morphology, supervised)
Arabic Segmentation (Habash and Sadat, 2006)

Problem Arabic fuses rich morphemes and clitics, often in complex ways

Example Alrys = Al+rys (the president)
wsynhY = w+s+y+nhY (and will finish (+ inflection))

Outline

1. Run a morphological analyser over the source-side
2. Split-off clitics and affixes
3. Train a standard SMT model

Example 2 (source-side morphology, supervised)
Arabic Segmentation (Habash and Sadat, 2006)

- Six splitting schemes, including "English-like" (**EN**)
- Also: regular expressions vs morphological analyser vs disambiguating morphological analyser
- Also: vary amount of training data: 1%, 10%, 100% (50k words to 5M)
- Compare BLEU for all combinations
- **EN** best with 1% corpus, but comparatively poor with 100%

Input	wsynhY
Gloss	and will finish

ST	wsynhY
D1	w+ synhy
D2	w+ s+ ynhy
D3	w+ s+ ynhy
MR	w+ s+ y+ nhy
EN	w+ s+ >nhY$_{VBP}$ +S$_{3MS}$

Unsupervised Segmentation?

- For segmentation, unsupervised models have recently been shown to rival supervised models
  - Clifton and Sarkar (2011) achieve best published result on English-Finnish task using an unsupervised morphological segmenter with a supervised post-processing merge step
  - Stallard et al. (2012) compare various supervised and unsupervised approaches on segmentation of Arabic (translation into English)

Baseline	43.5	
Best supervised	**46.5**	← highly engineered
Next best	45.6	
Unsupervised	45.8	

Example 3 (target-side morphology, supervised)
Russian Inflection (Toutanova et al., 2008)

Problem Russian has a rich system of inflection with many distinct forms for each lexeme

Example **си́ний** = blue (nom,sg,masc)

си́ними = blue (ins,pl)

Outline

1. Learn a model to predict inflection from stemmed MT output
2. Either a) Stem the target side of the training data, or b) Stem the output
3. Train a standard SMT model
4. Post-process: use the model to inflect the translations

Example 3 (target-side morphology, supervised)
Russian Inflection (Toutanova et al., 2008)

- Word represented as stem + vector of seven features
- MaxEnt Markov model predicts most likely inflected form from source-side and preceding words

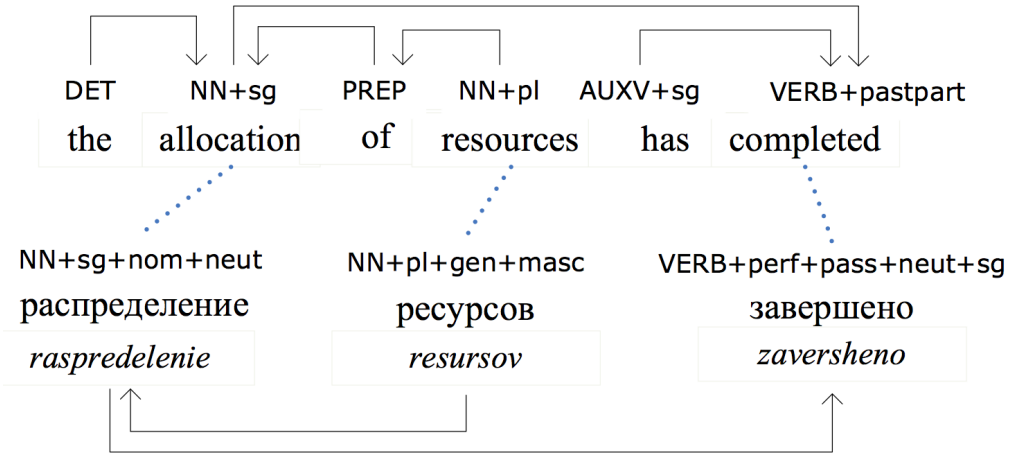


Figure 1: Aligned English-Russian sentence pair with syntactic and morphological annotation.

- Compare i) re-inflection with ii) stem inflection. Latter is 1 BLEU point better and 2 BLEU points above baseline.
- Improvement with ii) partly attributable to word alignment.

Summary for Pre-processing / Post-processing

- Subdivides translation problem
- But hard to predict what will work in MT
- And efficacy is dependent on training data size

Outline

Example 1 (target-side morphology, supervised)
Discriminative Lexicon Model (Jeong et al., 2010)

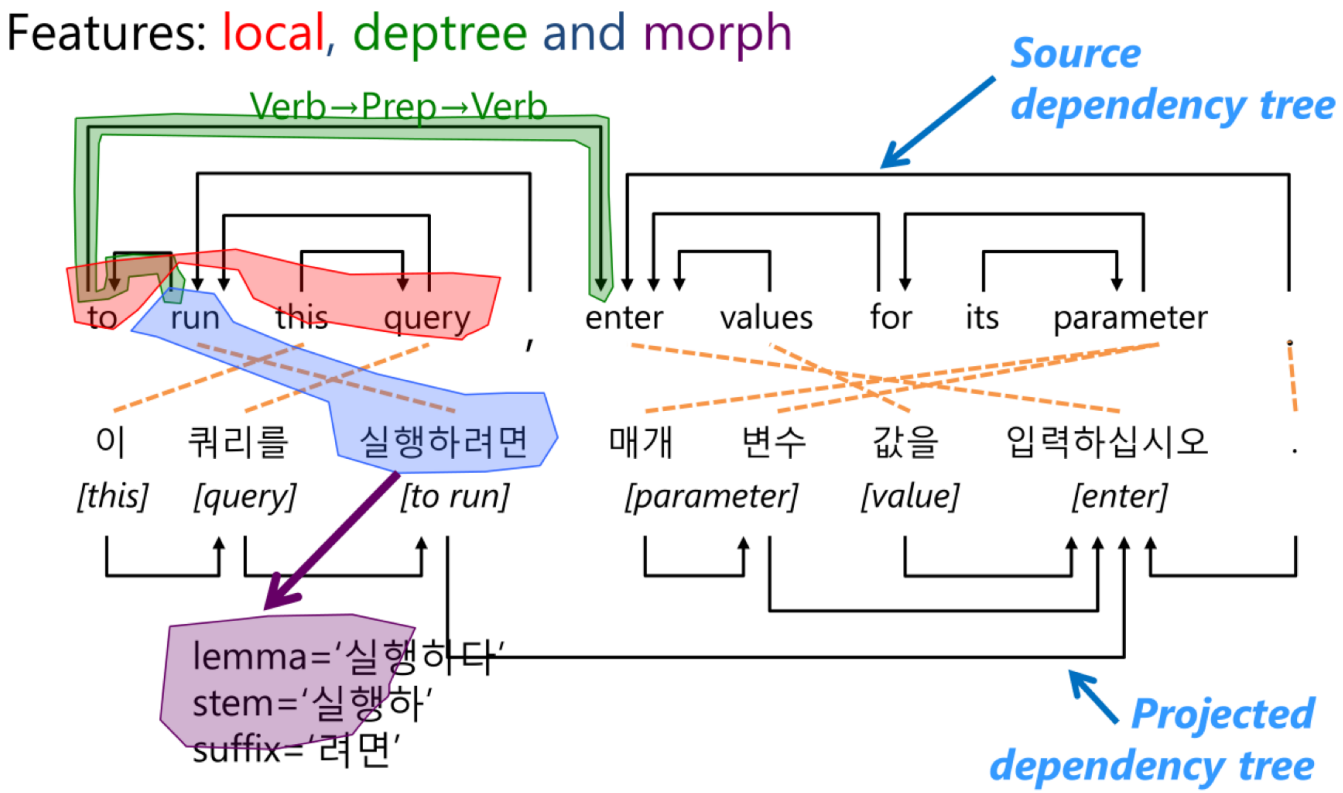
Goal Better selection of translation units by taking morphological (and other) information into account

Applicability Rich target-side inflection

Outline

1. Train a discriminative model to score target word according to window of surrounding words, aligned source word, and full source sentence
2. Add two feature functions: i) log-probability of target word according to model ii) indicator if target not observed with aligned source word during training

Example 1 (target-side morphology, supervised)
Discriminative Lexicon Model (Jeong et al., 2010)



- Log-linear model estimates probability of target word
- Gains of 0.6 / 0.5 / 0.5 BLEU on Bulgarian, Czech, Korean

Example 2 (target-side morphology, supervised)
Agreement and Segmentation (Green and DeNero, 2012)

Goal Better agreement across target phrase boundaries

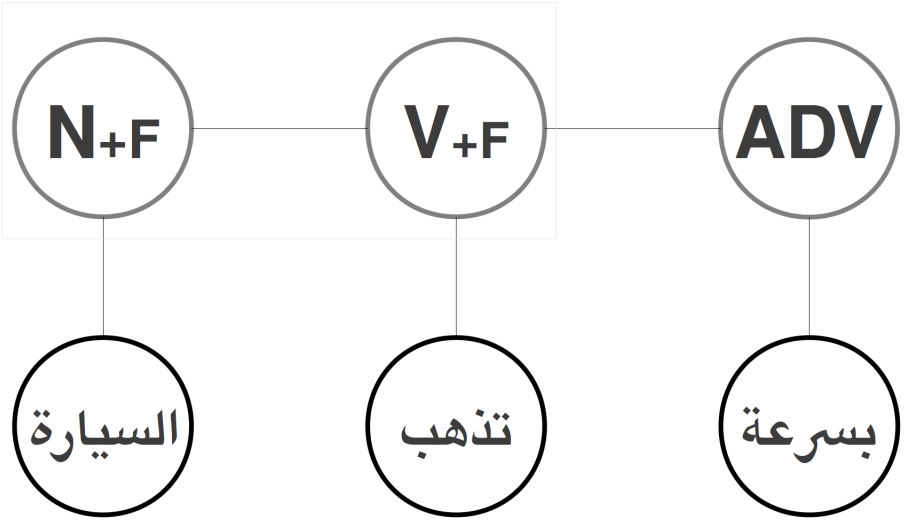
Applicability Local agreement

Outline

1. Train segmenter on Penn Arabic Treebank (ATB)
2. Train fine-grained sequence tagger on ATB
3. Add feature function that segments and tags hypotheses

Example 2 (target-side morphology, supervised)
Agreement and Segmentation (Green and DeNero, 2012)

- Scoring model performs its own segmentation (character-level CRF) then tags segments
- Outputs score of tag sequence
- Bi-gram: decoder must store previous segment and tag
- 1 BLEU point gain on large training data set with strong LM



Scoring in the Decoder?

Does integrating morphological knowledge into the decoding process help?

- It can do better than standard baseline
- But little (if any) empirical comparison between pre-/post-processing and integrated approaches on same task

Outline

Example 1 (source-side morphology, supervised) Confusion Network Input (Dyer, 2007)

Goal Allow decoder to back-off to simplified word forms

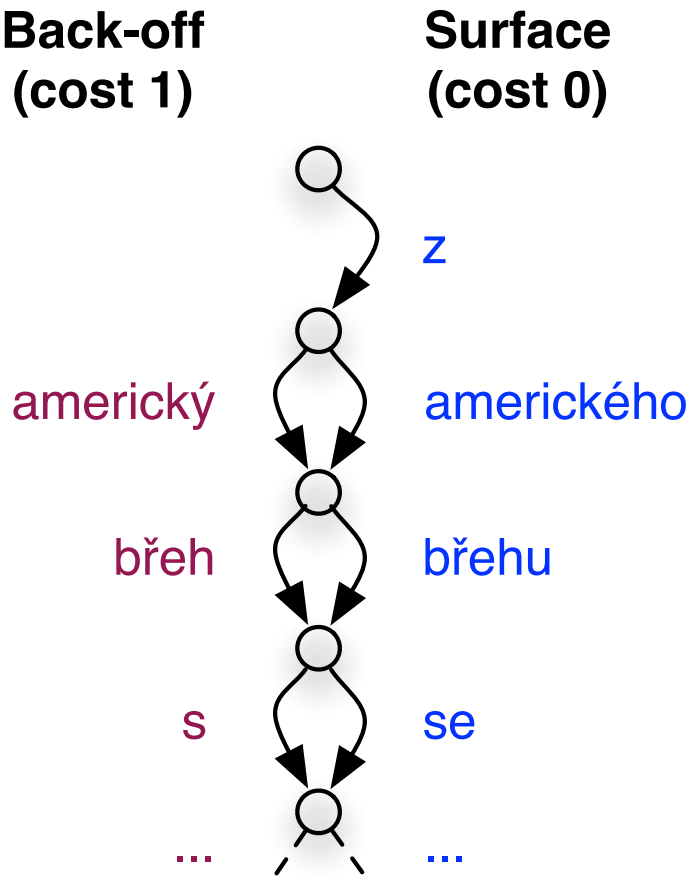
Applicability Rich source-side inflection (but see Dyer et al. (2008) for generalization)

Outline

1. Simplify source-side: $S \rightarrow S'$
2. Extract translation grammars from S and S'
3. Merge grammars and renormalize scores
4. Construct confusion network (CN) for input
5. Decode CN using feature function for back-off penalty

Example 1 (source-side morphology, supervised)
Confusion Network Input (Dyer, 2007)

- Experiments performed on Czech-English
- Compares lemmatization and truncation (lemmas work best)
- CN model outperforms pure pre-processing approach
- Approach later generalized to lattice input (Dyer et al., 2008)



Example 2 (source and/or target, typically supervised)
Factored Translation Models (Koehn and Hoang, 2007)

Goal Model words at multiple linguistic levels rather than as atomic tokens

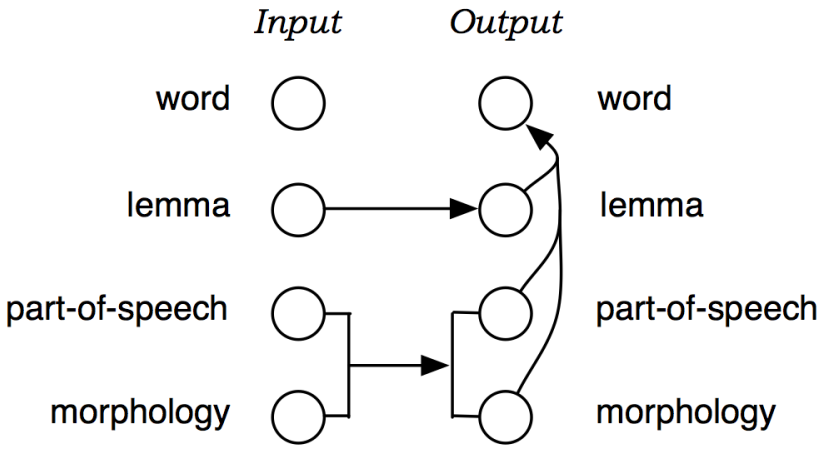
Applicability Source and target inflection. Harder to apply to, e.g., compounding but has been done.

Outline

1. Define source and target factorizations, e.g. (surface, POS, morph) and mapping
2. Factorize training and test data
3. Learn factor mapping and generation models from data
4. Decoder constructs translation options according to factored model

Example 2 (source and/or target, typically supervised)
Factored Translation Models (Koehn and Hoang, 2007)

- Extensive research
- Sequence models over factors can be higher-order than surface LM: e.g. 7-gram POS model is typical
- Some kinds of factorization introduce computational problems due to combinatorial explosion of translation options



Example 3 (target-side morphology, supervised)
Unification-based Constraints (Williams and Koehn, 2011)

Goal Improve agreement in morphologically rich target-side

Applicability Target-side agreement

Outline

1. Learn lexicon of agreement features for target words
2. Annotate target-side phrase structure trees with agreement relations
3. Extend SCFG rule extraction to include agreement constraints
4. Apply constraints during decoding, reject or downweight hypotheses on failure

Example 3 (target-side morphology, supervised)
Unification-based Constraints (Williams and Koehn, 2011)

- Feature values from morphological analyser
- Hand-written tool for annotation of trees with agreement relations
- Compared with factored translation, removes problem of combinatorial explosion:
  - Removes problem of combinatorial explosion
  - But can't generate novel forms

$$\textit{Kätzchen} \rightarrow \begin{bmatrix} \text{CAT} & \texttt{NN} \\ \text{INFL} & \begin{bmatrix} \text{AGR} & \begin{bmatrix} \text{GEN} & \texttt{n} \\ \text{NUM} & \texttt{sg} \\ \text{PER} & \texttt{3} \end{bmatrix} \\ \text{CASE} & \texttt{nom} \end{bmatrix} \end{bmatrix}$$

$$\textit{das} \rightarrow \begin{bmatrix} \text{CAT} & \texttt{ART} \\ \text{INFL} & \begin{bmatrix} \text{AGR} & \begin{bmatrix} \text{GEN} & \texttt{n} \\ \text{NUM} & \texttt{sg} \\ \text{PER} & \texttt{3} \end{bmatrix} \\ \text{CASE} & \texttt{nom} \\ \text{DECL} & \texttt{weak} \end{bmatrix} \end{bmatrix}$$

$$\text{NP-SB} \rightarrow \text{X}_1\ \textit{kitten} \mid \text{ART}_1\ \textit{Kätzchen}$$
$$\langle \text{ART INFL} \rangle = \langle \textit{Kätzchen}\ \text{INFL} \rangle$$

Outline

Conclusion

Modelling morphology in MT:

- Varies in terms of subtasks types and integration into decoding
- Can be highly effective.
- But is still a bit of a black art:
  - There are no one-size-fits-all solutions.
  - Performance on subtasks doesn't always predict effectiveness for MT.
  - Effectiveness depends on training data size.
  - Little research into controlled comparison
- Traditionally relied on supervised processing, but that's starting to change.

References

- *Empirical Methods for Compound Splitting*
  Philipp Koehn and Kevin Knight. EACL 2003.
- *Europarl: A Parallel Corpus for Statistical Machine Translation*
  Philipp Koehn. MT Summit 2005.
- *Arabic Preprocessing Schemes for Statistical Machine Translation*
  Nizar Habash and Fatiha Sadat. NAACL 2006.
- *Factored Translation Models*
  Philipp Koehn and Hieu Hoang. In EMNLP 2007.
- *The 'Noisier Channel': Translation from Morphologically Complex Languages*
  Chris Dyer. WMT 2007.
- *Generalizing Word Lattice Translation*
  Chris Dyer, Smaranda Muresan, Philip Resnik. ACL 2008.
- *Applying Morphology Generation Models to Machine Translation*
  Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. ACL 2008.

References

- *A Discriminative Lexicon Model for Complex Morphology*
  Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. AMTA 2010.
- *Agreement Constraints for Statistical Machine Translation into German*
  Philip Williams and Philipp Koehn. WMT 2011.
- *Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction*
  Ann Clifton and Anoop Sarkar. ACL 2011.
- *A Class-Based Agreement Model for Generating Accurately Inflected Translations*
  Spence Green and John DeNero. ACL 2012.
- *Unsupervised Morphology Rivals Supervised Morphology for Arabic MT*
  David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, Regina Barzilay. ACL 2012.