Using TectoMT as a Preprocessing Tool for Phrase-Based SMT

Daniel Zeman

ÚFAL MFF

Univerzita Karlova v Praze

Charles University in Prague

The research has been supported by the grant MSM0021620838.

Outline

- Phrase-based statistical machine translation
- TectoMT
- Preprocessing for MT
- Overview and motivation of transformations
- Preliminary results

Phrase-Based Statistical Machine Translation

- Sentence-aligned bilingual parallel corpus
- Automatically compute *(estimate)* word alignment
- Based on word alignment, find possible *parallel phrases* (sequences of words)
- In *hierarchical systems* (Chiang 2005), phrases may contain gaps (non-terminals)
- We use Joshua, an open-source hierarchical system
  - http://sourceforge.net/projects/joshua/

Phrase-Based Statistical Machine Translation

- Target language model
- Translation hypotheses are scored according to
  - Translation model (3 scores)
  - Target language model (1 score)
- Minimum Error Rate Training (MERT)
  - Tunes the weights of the various scores (features) on held-out data
  - Must be able to automatically judge translation quality
    ⇒ BLEU score

TectoMT

- TectoMT is a system for machine translation
- Unlike Joshua, this is *not a phrase-based system*
- It is not even *statistical MT* in the usual sense
  - But it contains many statistical components anyway: taggers, parsers, word frequency lists etc.

- TectoMT is based on the traditional pyramid-like paradigm: analysis of the source language – transfer – synthesis of the target language
- http://ufal.mff.cuni.cz/tectomt/ (licensed under GPL)

TectoMT

- TectoMT is highly modular
- Dozens of blocks of code (in Perl) are applied to the same text, one after the other
- TectoMT provides common interface to the textual data:
  - token = node (of a tree)
    - token attributes, e.g. *lemma, morpho-tag, dependency-label…*
  - nodes are organized in trees
  - easy tree manipulation (`get_children()`, `set_parent()`, `shift_after_node()`…)

TectoMT

- Some code-blocks are rather tiny, e.g.
  - Search for punctuation nodes, normalize *“fancy quote marks”* to *``Penn Treebank style''*
- Others may be long and complex, e.g.
  - Look for all personal pronouns, find the probable noun phrase they refer to, store the link for later blocks that will check whether translation changed the gender
    - en: *a bag lay on it [the chair]* … neuter
    - cs: *na ní [židli] ležela taška* … feminine
- Yet others encapsulate calls to external software
  - Taggers, parsers, named entity recognizers…

TectoMT

- All blocks work with common interface and common data format
- Easy to modify your *scenario* by e.g.
  - unplugging the block with Collins parser
  - replacing it by a block with Stanford parser
- The framework is language-independent but many blocks must obviously be language-specific

- Existing scenarios (block sequences) are ready to reuse, especially for the analysis of English and Czech

TectoMT as a Preprocessor

- TectoMT is not just an MT system
- It is an NLP framework useful for various purposes

- Out of the analysis – transfer – synthesis sequence, we use only some of the analysis blocks
- We implement *new blocks* that operate on dependency trees and *transform* them
  - Change nodes (word forms)
  - Insert or remove nodes
  - Reorder nodes

TectoMT as a Preprocessor

- After analysis and transformation, we use a `Print` block to extract plain text from the TectoMT data structures

- The transformed plain text is used as a new training corpus for Joshua (the statistical MT system)

- Motivation: well aimed transformations of the training data could make learning of parallel phrases easier

SMT and Preprocessing

- There is a body of previous related work
    - Nießen & Ney (2004)
    - Collins et al. (2005)
    - Popović et al. (2005)
    - Goldwater & McClosky (2005)
    - Habash & Sadat (2006)
    - El Isbihani et al. (2006)
    - Prokopová (2007)
    - Avramidis & Koehn (2008)
    - Axelrod et al. (2008)
    - Popović et al. (2009)
    - Ramanathan et al. (2009)

Related Work

- Nießen & Ney (2004): de-en: compound splitting, separable verb prefixes rejoin verbs
- Collins et al. (2005): de-en: source text parsing, then reordering transformations
- Popović et al. (2005): sr-en: lemmatization, verb person → personal pronoun; en-sr: removal of articles
- Goldwater & McClosky (2005): cs-en: lemmatization, then partial restoring of morphology
- Habash & Sadat (2006), El Isbihani et al. (2006): ar-en: retokenization of Arabic

Related Work

- Prokopová (2007): cs-en: reordering, inserting (into Czech) *to, of, by*
- Avramidis & Koehn (2008): en-el: acquire English syntactic functions ⇒ generate Greek case markers
- Axelrod et al. (2008): de-es: German stemming and compound splitting
- Popović et al. (2009): de-en, fr-en, es-en: part-of-speech-based source reordering
- Ramanathan et al. (2009): en-hi: reordering (SVO to SOV); English syntactic functions ⇒ Hindi suffixes

Preprocessing Source Only

- We can preprocess the source side of
  - training data
  - development and test data
- We don’t touch the target side!
  - Can’t preprocess target test data — the system must generate it
  - Preprocessing the reference translation would be cheating
- Theoretically, we could
  - Preprocess training data and
  - Postprocess the system output for test data (reverse transformation)
  - More difficult (the system output may be ungrammatical)

Our Work

- Source language is English
  - Multitude of available tools
  - We use standard TectoMT pipeline for English analysis:
    - Morče tagger (http://ufal.mff.cuni.cz/morce/)
    - MST dependency parser (http://sourceforge.net/projects/mstparser/)
    - ~ 40 other code blocks
- Two typologically different target languages for comparison:
  - Czech (obvious reasons)
  - Hindi (NLP Tools Contest)

Possible Transformations

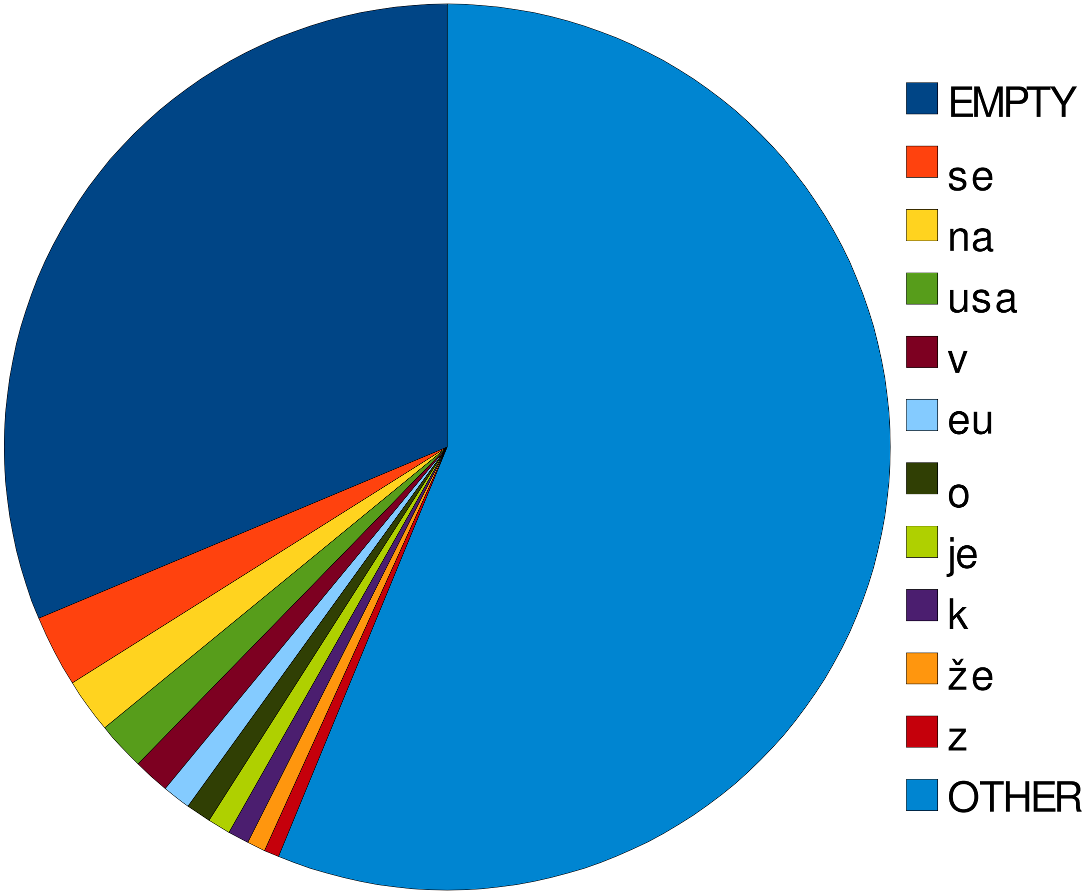
- en-cs
  - Remove articles
  - Target case selection
  - (Target agreement)
  - Verbal groups
  - Personal pronouns
  - *and more…*

- en-hi
  - Remove definite articles
  - Target case selection
  - (Target agreement)
  - Change prepositions to postpositions
  - Subject-object-verb order
  - The verb *to have*
  - *and more…*

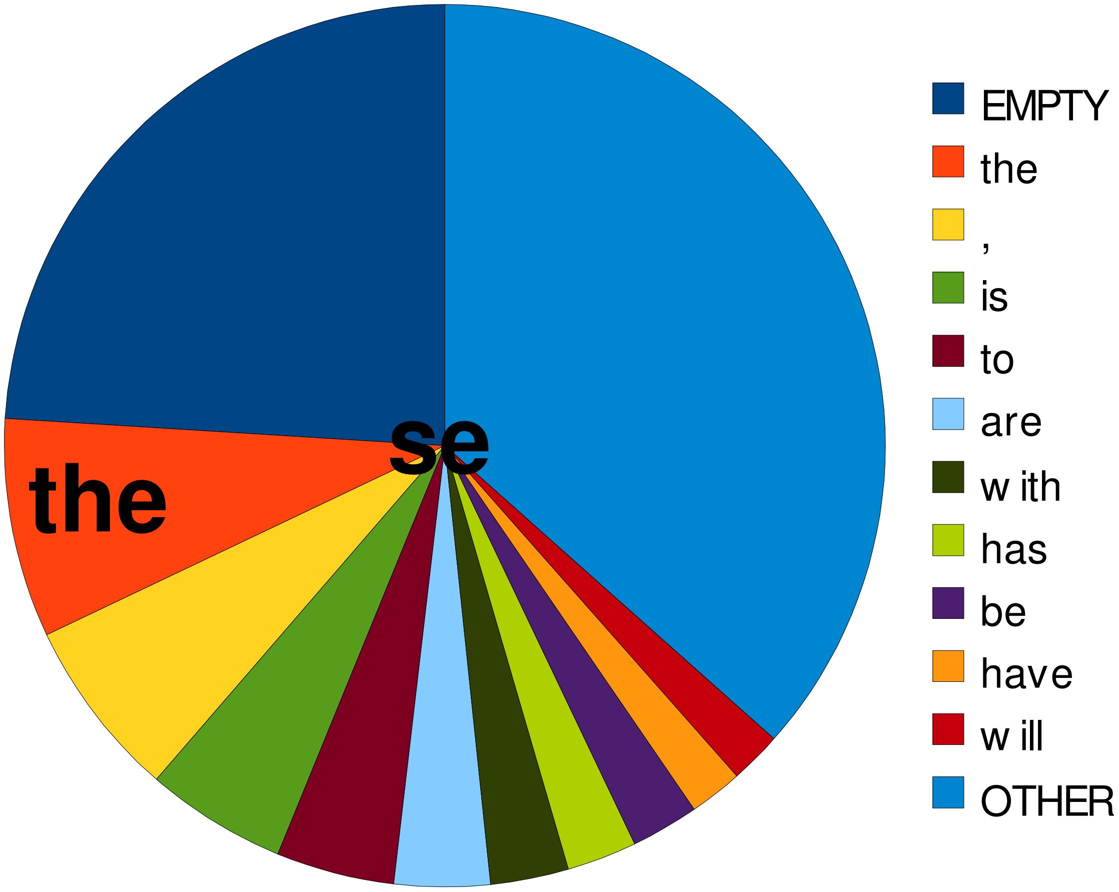
Remove English Articles

- No articles in Czech
- Word aligner might (correctly) decide that *the* corresponds to empty word
- However, quite often it will align to neighboring words
- Unnecessarily increases data sparseness:
  - cs: *pražskou*
  - en:
    - *the Prague*
    - *Prague the*

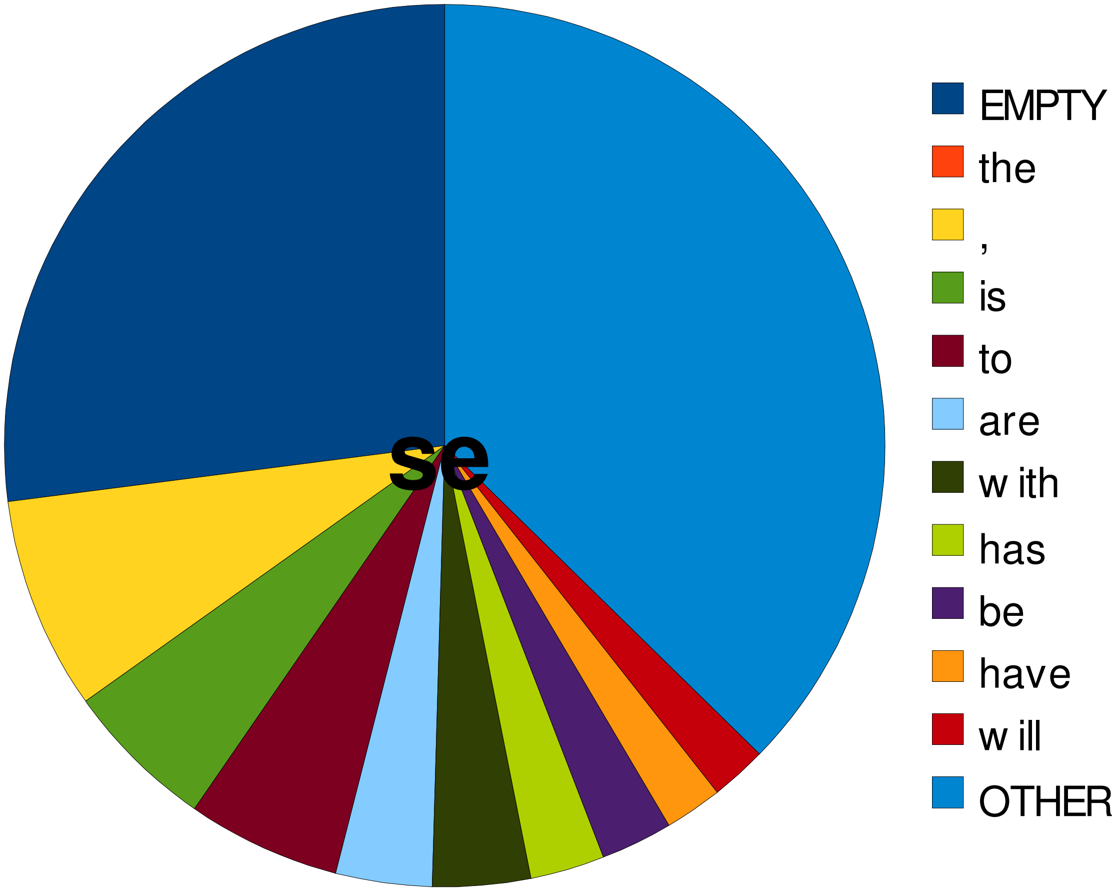
Czech Alignments of *the*



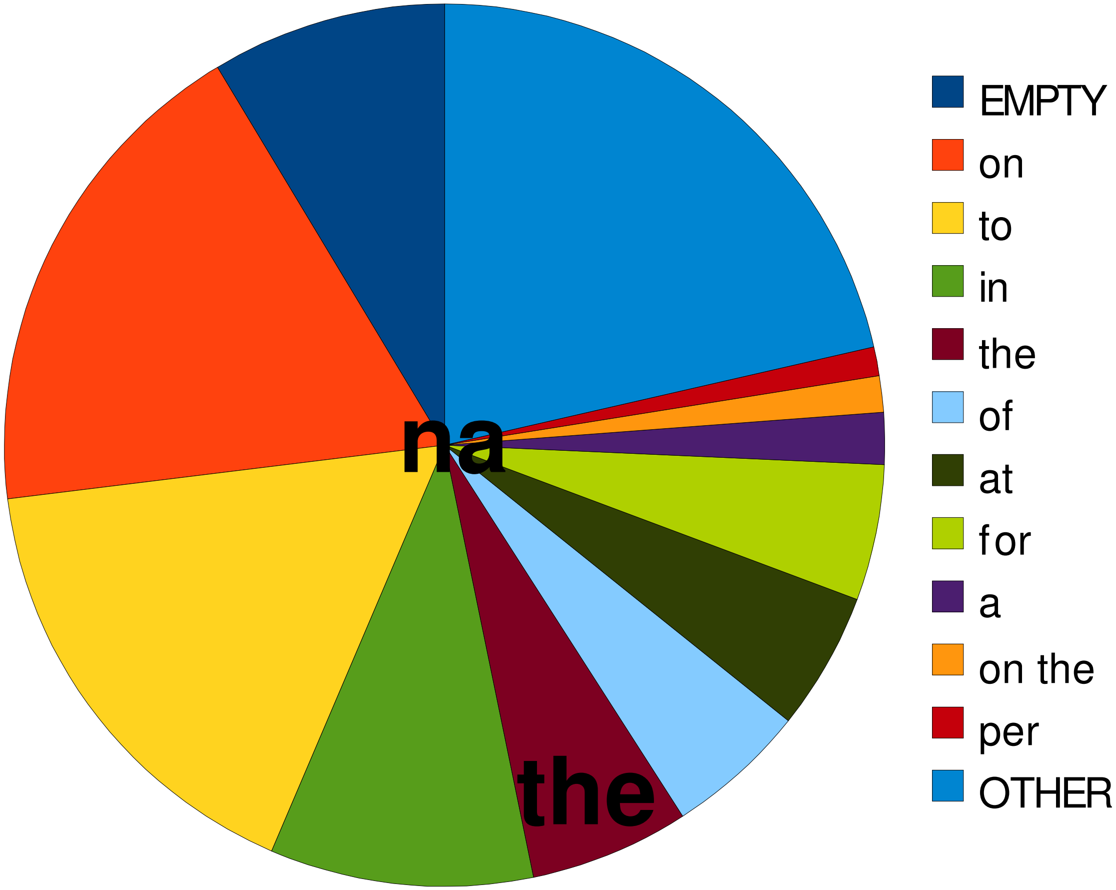
Alignments of *se*



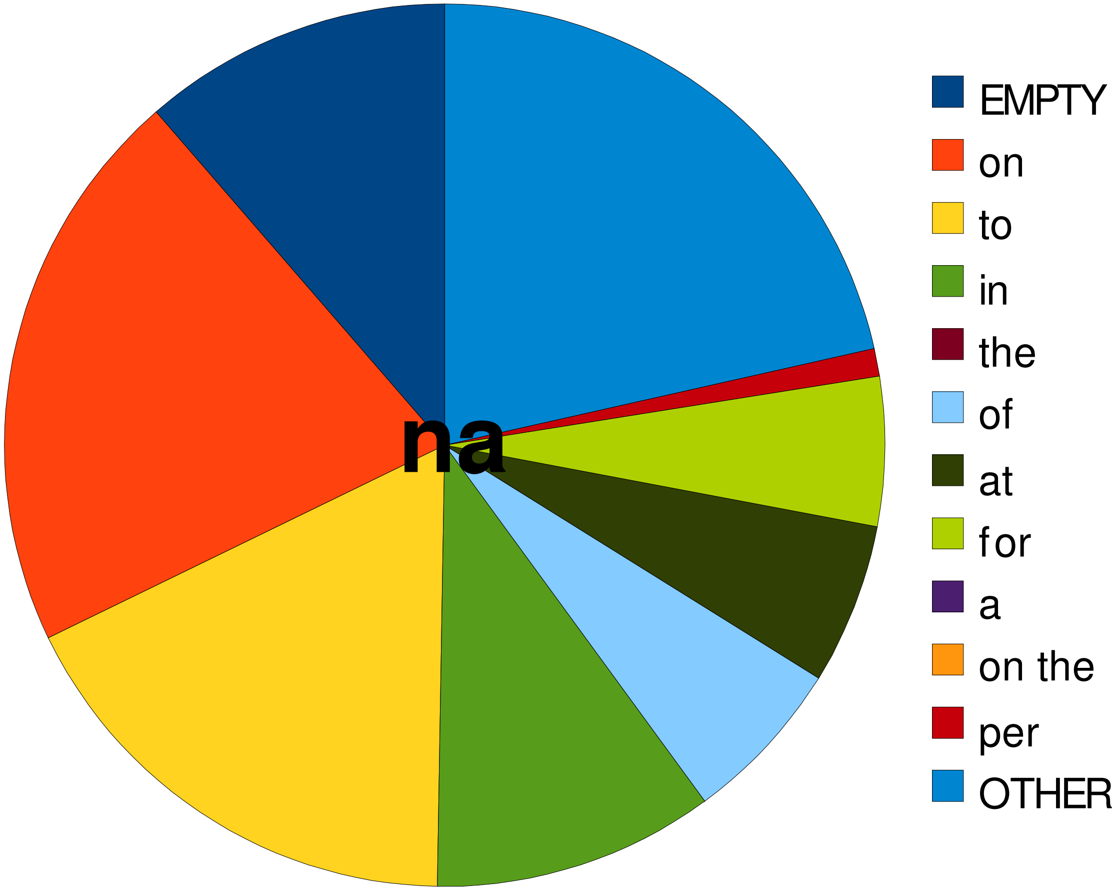
Alignments of *se* after removing articles



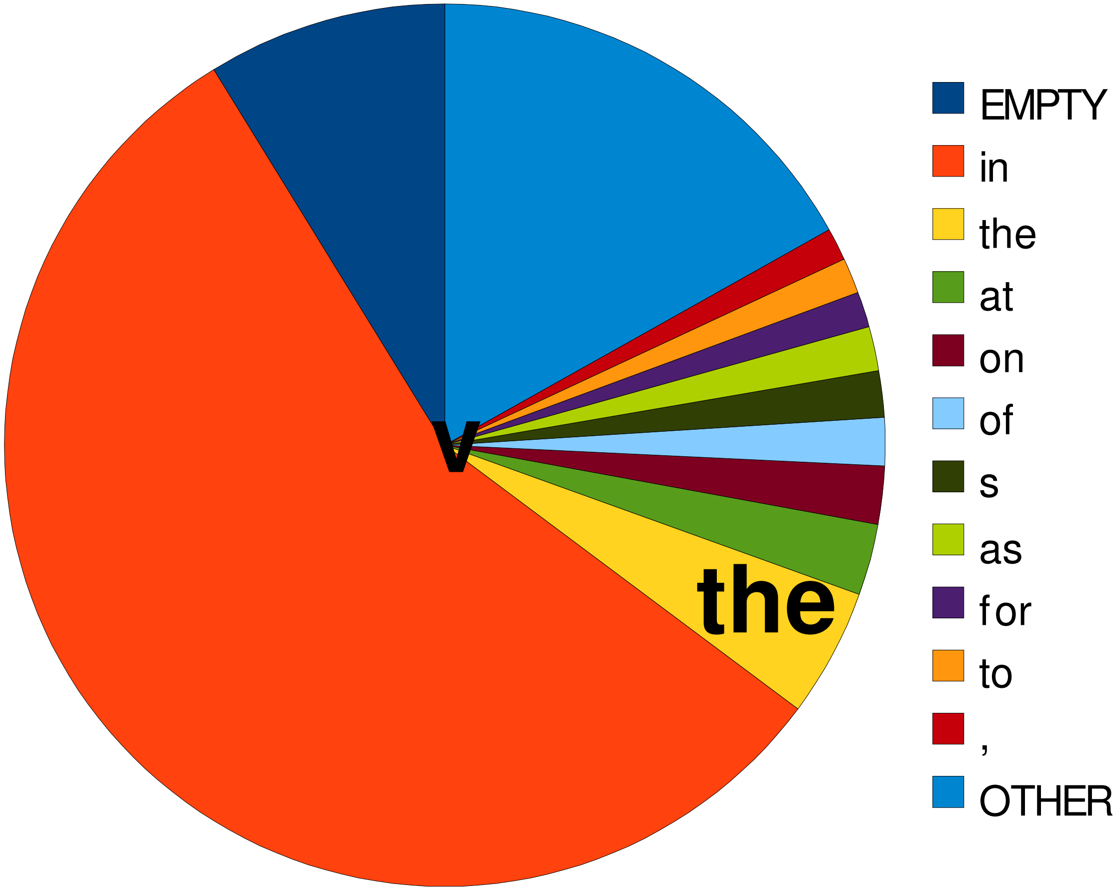
Alignments of *na*



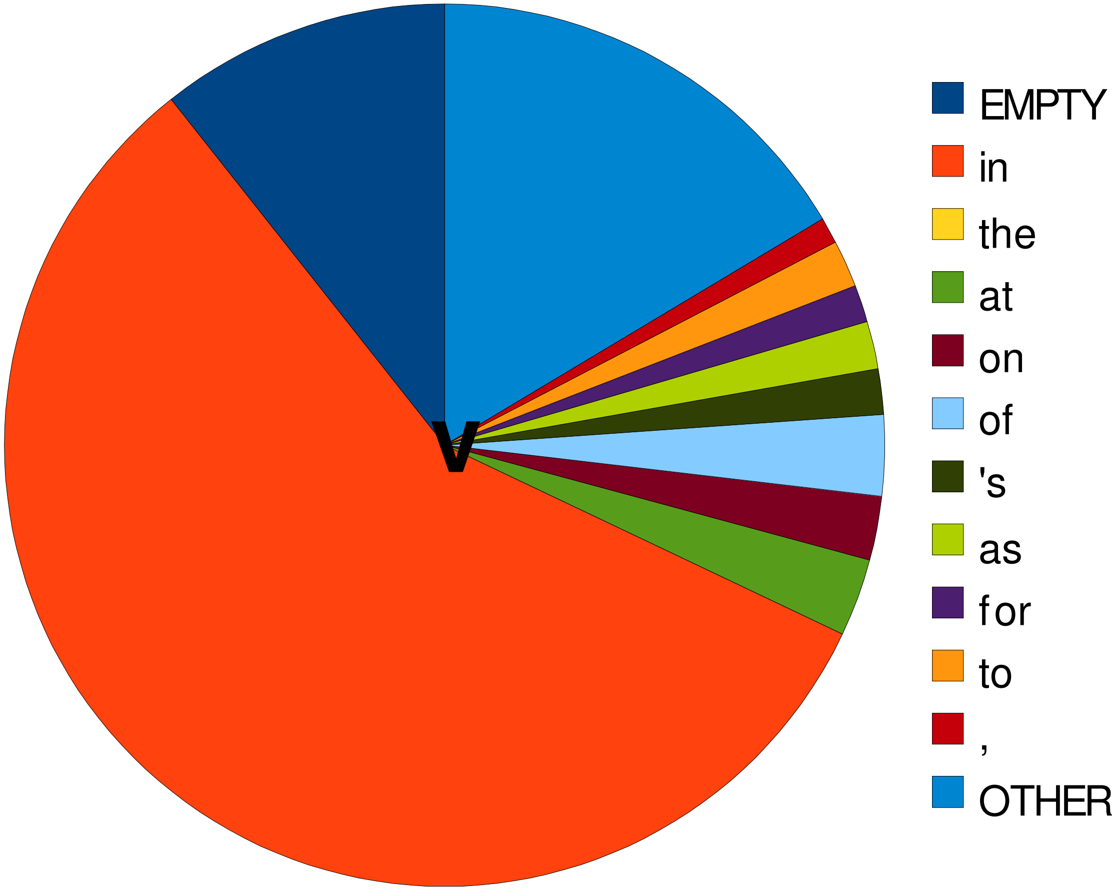
Alignments of *na* after removing articles



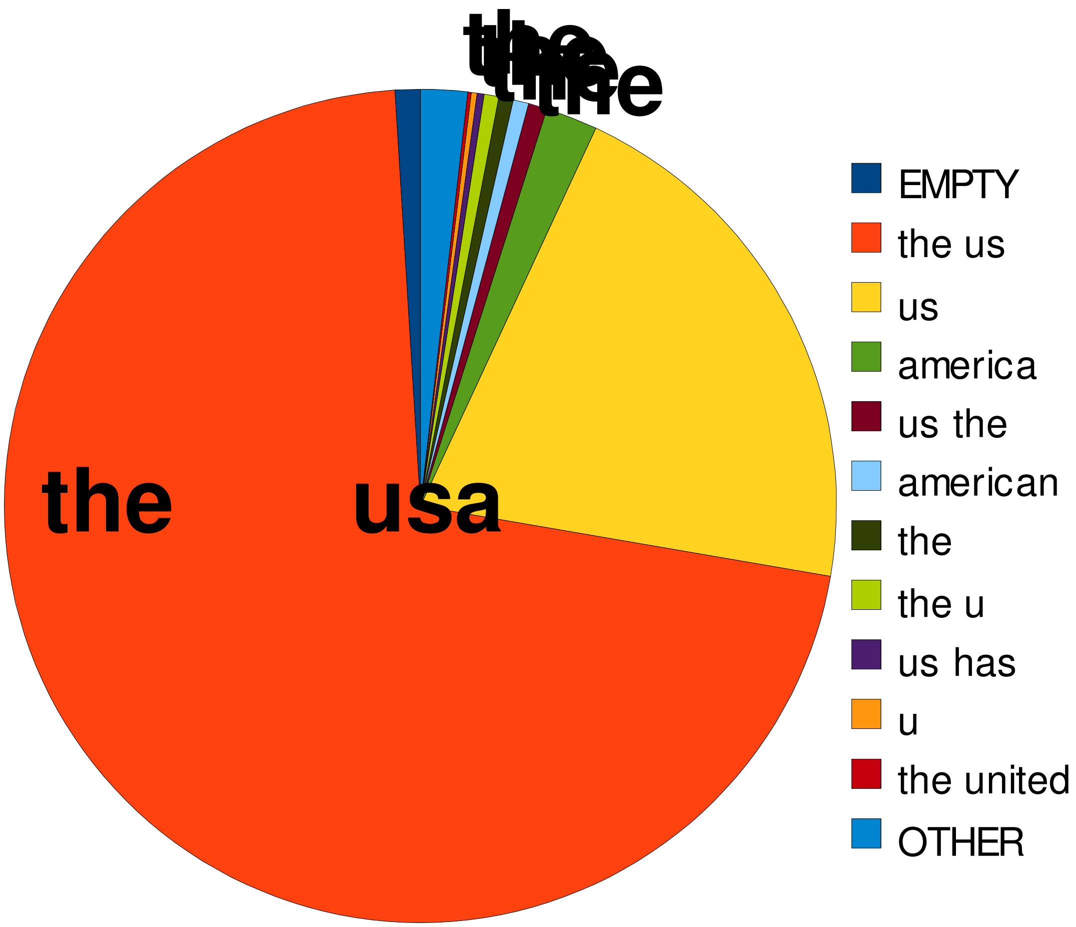
Alignments of *v*



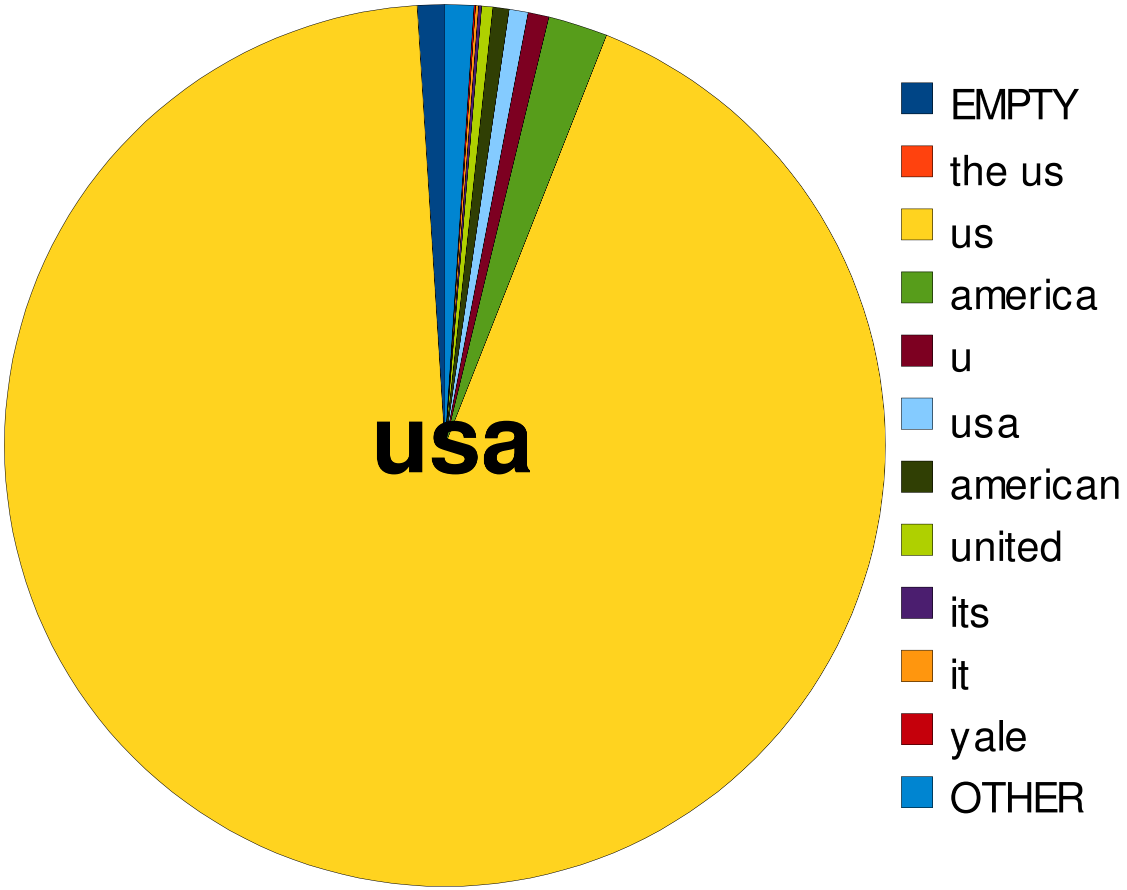
Alignments of *v* after removing articles



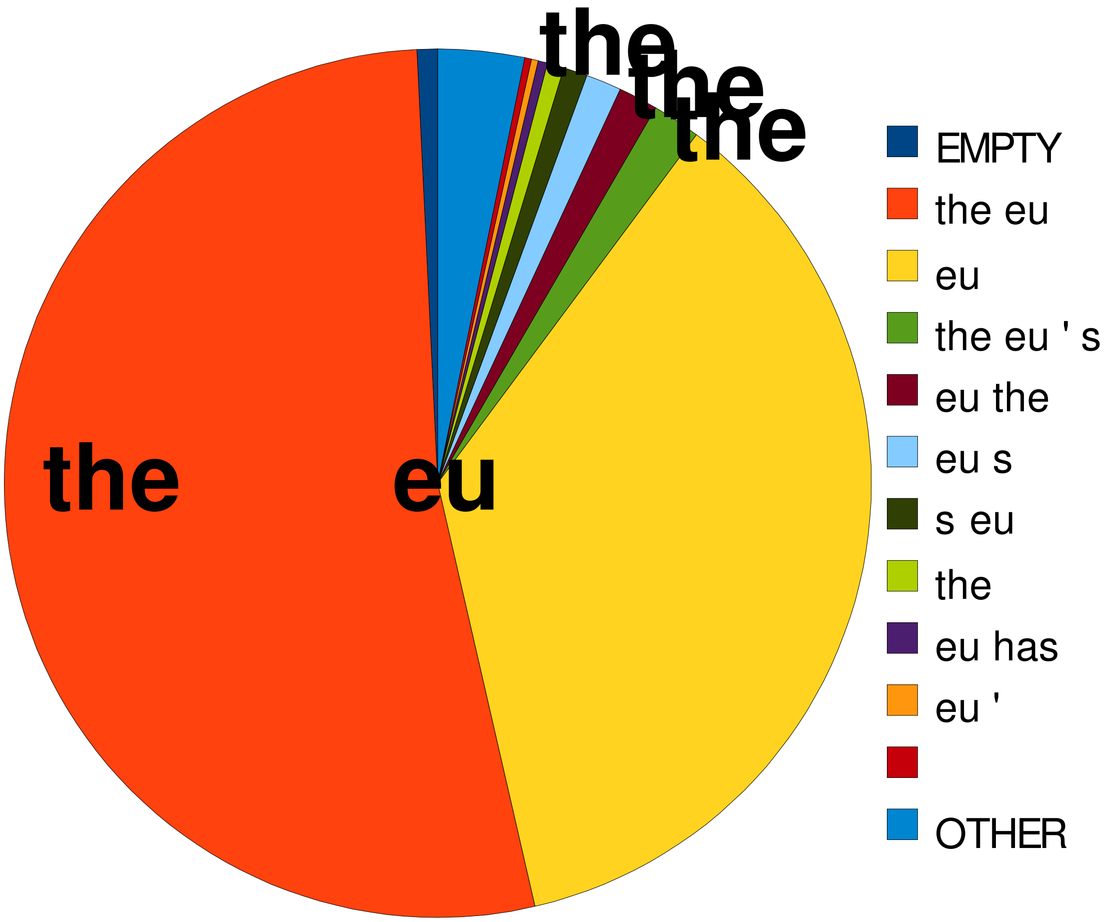
Alignments of *usa*



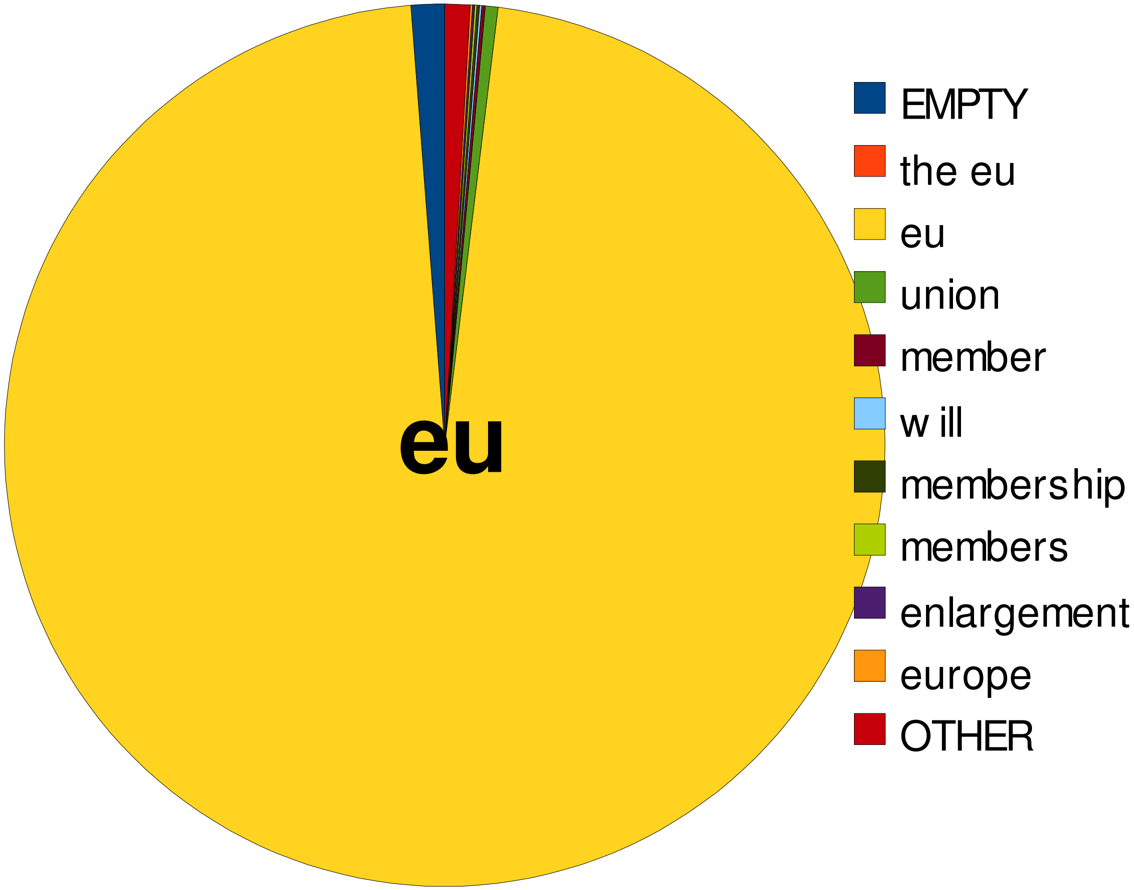
Alignments of *usa* after removing articles



Alignments of *eu*



Alignments of *eu* after removing articles



Target Case Selection

- Almost no case marking in English
  - 7 cases in Czech
  - 2 cases / ~8 *vibhakti* in Hindi
- We cannot preprocess the target side
- However, we can explicitly mark syntactic functions

- Hopefully the system will learn that
  - *mother_Sb* → *matka* (nom.) | *माँ ने (mã̃ ne)* (agent.)
  - *mother* → (other cases)

Verbal Groups

- Complex system of tenses and aspects in English
- Czech is simpler
- All English auxiliaries should be close to the main verb
  - Otherwise, higher risk that they will be translated separately

- *he is now finally coming* → *he comes now finally*
  - No continuous tenses in Czech
- *he has never achieved* → *he achieved never*
  - Only simple past in Czech

Personal Pronouns

- Czech is a pro-drop language
  - Subject may be missing
  - Personal pronoun is not obligatory in that case
  - Finite verbs are marked for person and number
- As a result, English pronouns often lack counterparts
  - They should be aligned to Czech finite verbs
    - Sometimes they are, sometimes not
- Possible solutions:
  - Merge pronouns with their verbs such as *we-work*
  - Or at least make sure they are adjacent: *he always comes* → *always he comes*

Postpositions in Hindi

- English uses prepositions, Hindi postpositions
  - *घर में (ghara meñ) = house in*
  - *मेरे अध्यापक की किताब (mere adhyāpaka kī kitāba) = my teacher of book* = “my teacher’s book”
  - *राम की तरफ़ (rāma kī tarafa) = Ram of direction* = “towards Ram”

- Proposed transformation:
  - Move prepositions after their noun phrases
  - Transform patterns of the *X of Y* type to *Y of X*

Subject-Object-Verb Order

- Although the Hindi word order is said to be not as fixed as in English, verbs are usually found at the end
  - *एक मित्र के साथ कुछ काम कर रहा हूँ*
  - *eka mitra ke sātha kucha kāma kara rahā hūṁ*
  - *one friend of with some work do -ing am*
  - *I’m doing some work with a friend.*

- Proposed transformation:
  - Move finite verbs to the end of the subtree they dominate
  - Avoid skipping nested clauses

The Verb *to have*

- Similarly to Russian, Hindi has no direct translation of *to have*. Periphrastic constructions are used to convey the sense of having:
  - *हमारे पास समय नहीं है ।*
  - *hamāre pāsa samaya nahīṁ hai.*
  - *our at time not is.*
  - *We don’t have time.*

- Possible solution:
  - Make *to have* an exception to the verb reordering rule. Keep it with its subject and learn *X has* → *X के पास*

Preliminary Results

- So far we have tried
  - For en-cs:
    article removal, subject marking and verb tense simplification
  - For en-hi:
    article removal, postpositions and SOV reordering

- In terms of BLEU score, the results are not convincing (statistically insignificant change)
  - en-cs: 0.0863 → 0.0905
  - en-hi: 0.1006 → 0.1029

Preliminary Results

- Human inspection of the data suggests that the targeted phenomena are improving (e.g. the alignments of *the*)
- No large-scale human evaluation available yet
- Open questions:
  - How frequently do transformations apply, i.e. what is their potential to change translation results?
  - To what extent is the hierarchical system actually able to learn the reordering, even with the bad alignment?
  - How serious is the role played by tagging and parsing errors?

Example of a Parsing Error

- < the potential charges are serious : conspiring to destabilize the government that was elected last february , unlawfully removing the country ' s top judges in november 2007 , and failing to provide adequate security to benazir bhutto before her assassination last december .
- ---
- > the potential charges conspire serious : to destabilize the government that was elected last february , unlawfully removing the country ' s top judges in november 2007 , and failing to provide adequate security to benazir bhutto before her assassination last december .

Conclusion

- Showed how TectoMT can be used to easily implement various transformations of data for SMT
- Discussed translation from English to two different Indo-European languages, motivated and proposed a number of transformations
- Preliminary BLEU score results are not convincing
- Detailed human analysis is needed
  - Future research should also investigate postprocessing of the target side (rich morphology)

Thank you

Děkuji

धन्यवाद

The research has been supported by the grant MSM0021620838.