MonoTrans: Statistical Machine Translation from Monolingual Data

Rudolf Rosa <rosa@ufal.mff.cuni.cz>, Univerzita Karlova, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky, Praha, Česko

Setting

- similar languages (e.g. Slovak and Czech)
- no parallel data, no supervised source data

Goal

- simulate human-like crosslingual understanding
- perform machine translation without parallel data

Reasoning

- there are ~7000 languages in the world
- required data easy to get for ~100 languages

Method

Idea: translate each source word by the **most similar** target word

Assumptions: corresponding words:
- are string-wise similar
- have similar length
- have similar frequency

$$\mathrm{sim}(s,t) = \mathrm{sim}_{JW}(s,t) \cdot \mathrm{sim}_{len}(s,t) \cdot \mathrm{sim}_{freq}(s,t)$$

String-wise similarity

- based on Jaro-Winkler edit distance
- prefixes more important
  - good for languages with word-final inflection
- also computed on ASCII transliterations
- also computed on devowelled words (strip AEIOUY)

$$\mathrm{sim}_{JW}(s,t) = \mathrm{sim}_{jw}(s,t) \cdot \mathrm{sim}_{jwT}(s,t) \cdot \mathrm{sim}_{jwD}(s,t) \cdot \mathrm{sim}_{jwDT}(s,t)$$

Length similarity

- also computed on devowelled words

$$\mathrm{sim}_{len}(s,t) = \frac{1}{1+\left|\mathrm{len}(s)-\mathrm{len}(t)\right|}$$

Frequency similarity

- computed on available monolingual corpora
- normalization to account for differing corpora sizes

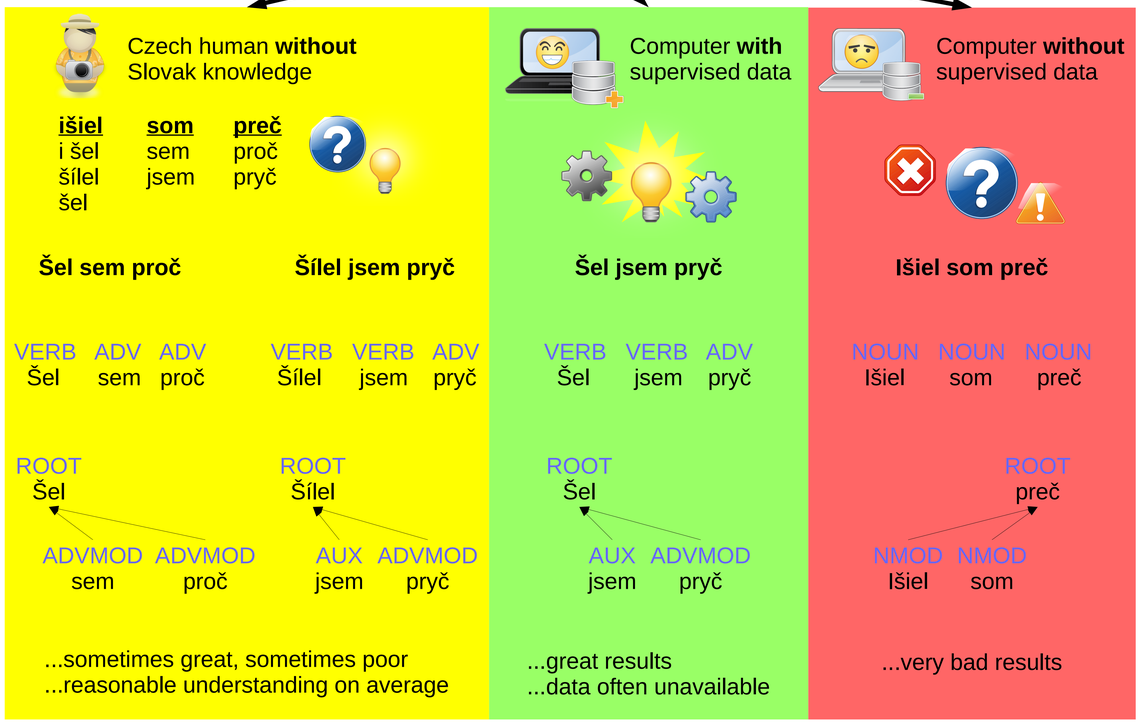
$$\mathrm{sim}_{freq}(s,t) = \frac{1}{1+\left|\log(\mathrm{freq}(s))-\log(\mathrm{freq}(t))\right|}$$

Efficiency

- cannot go over all target words for each source word
  - leads to e.g. 1 word/minute
  - goal: e.g. 1 word/second
- early stopping
  - start with frequent candidates
  - stop once sim_{freq} is lower than current highest sim
  - huge speedup
  - limited capabilities: hard to add a language model
- word list partitioning
  - source and target word must share first 2 consonants
  - huge speedup
  - introduces search errors, many translations not found
    - e.g. “som” cannot be translated to “jsem”

Motivation

Input: Išiel som preč



Sample outputs (SK-CS)

SK: Všetci sme počuli o tom, čo si urobil pre toho chlapca a je mi ľúto, že to nezvládol.
CS: Všech sem poučil o tom, čo si urobil pro toho chlapce a je mi líto, že to nezvládol.

SK: Navrhujem , aby si sa zase sústredil na hru.
CS: Navrhuji, aby si se zase sústredil na hru.

SK: Volá sa to vernosť.
CS: Volá se to věrnost.

SK: Keď ma zrazil opitý vodič, ktorého uznali nevinným, bol to Doug, kto mi pomohol zaplatiť za štúdium práva.
CS: Kde má zrazil opít vodil, kterého uznali nevinnou, byl to Doug, který mi pomohl zaplatit za štúdium práva.

SK: Posledné, čo sa mi chce, je strčiť ti loptu do ruky.
CS: Poslední, čo se mi chce, je strčte ti lopaty do ruky.

SK: Zoznámim vás s priateľmi.
CS: Zazvoním vás s parazitem.

SK: Počuj, asi som zaujatý, no toto je dokonalá svadba.
CS: Počkej, asi sem zajatí, no toto je dokonalý svatba.

SK: Môžem úprimne povedať, že som nikdy nemal priateľa, s ktorým by som sa cítil ako s tebou.
CS: Můžeme promine povídat, že sem nikdy neměl promiňte, s kterým by sem se cítil ako s tebou.

SK: Nie! Žiadne arašidové maslo!
CS: Ne! Žádné arašidové mysli!

SK: My sa môžeme povaľovať sa na pohovkách.
CS: My se můžeme považovat se na pohádkách.

SK: Na našom treťom rande povedala, že si chce založiť rodinu, že je pripravená mať deti.
CS: Na našim treťom rande povídal, že si chce založili rodina, že je připravená mít děti.

SK: Alebo povedať: Doug, poďme na pivo. Alebo si pozrime zápas.
CS: Alebo povídat: Doug, podle na pivo. Alebo si pozřeme zápas.

SK: Mohol by sa aj zvyšok svadobčanov pridať k šťastnému páru?
CS: Mohl by se já zvykem svadobčanov prodat k strašnému páru?

SK: Historická pánska jazda s mojím najbližším priateľom.
CS: Historický panské jízdě s mojí nejbližším parazitem.

SK: Nekecaj. Myslíš, že ja to nechcem?
CS: Nekecej. Myslíš, že já to nechceš?

SK: Vyhlasujem vás za manželov.
CS: Vyhlašujeme vás za manželku.

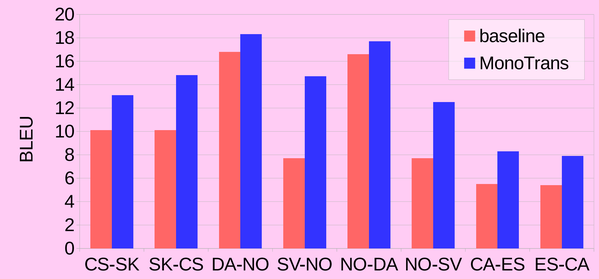
CS: Snažím sa bojovať so svojimi zbraňami každý deň, každý zápas, po celú dobu svojej kariéry.
SK: Snažím se bojovat s svojich zbraněmi každý den, každý zápas, po celé dobu svoje kariéry.

SK: Nemal by som jej vravieť pravdu?
CS: Neměl by sem její vrahovi pravdu?

Evaluation

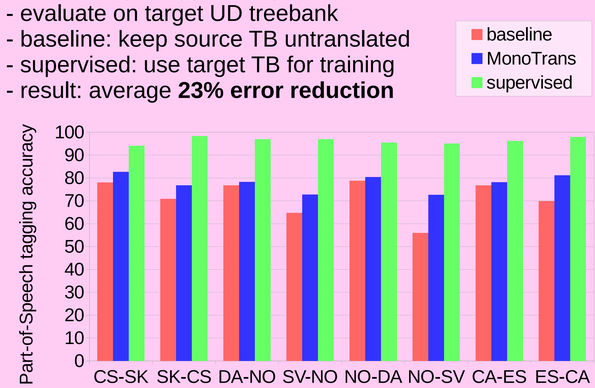
Intrinsic evaluation: translation quality

- data source: OpenSubtitles2016
  - Czech (CS), Slovak (SK)
  - Danish (DA), Norwegian (NO), Swedish (SV)
  - Catalan (CA), Spanish (ES)
- data size: 10,000 sentences
- baseline: keep source text untranslated
- result: average **43% improvement** over baseline



Extrinsic evaluation: tagging and parsing

- treebank data: Universal Dependencies v1.4
- translate source UD treebank into target language
- train UDPipe tagger and parser
- evaluate on target UD treebank
- baseline: keep source TB untranslated
- supervised: use target TB for training
- result: average **23% error reduction**



This work has been supported by the grant No. DG16P02B048 of the Ministry of Culture of the Czech Republic, the grant No. CZ.02.1.01/0.0/0.0/16_013/0001781 of the Ministry of Education, Youth and Sports of the Czech Republic, and the SVV 260 453 grant. This work has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).