

Hálek – Rosa – Tamchyna

Machine Translation
of Named Entities
with Help of Wikipedia

Named Entity Translation

- **Rice University** is at 6100 **Main Street**.
- **Steven Bird** passed on the editorship...
- **fork()** creates a new process.
- **Univerzita rýže** je v 6100 **hlavní ulici**.
- **Steven pták** přenesl na editorship...
- **vidlička()** vytváří nový proces.

Google Translate

Google překladač

Z: česky ▾



Do: anglicky ▾

Přeložit

Žiju v Plzni.

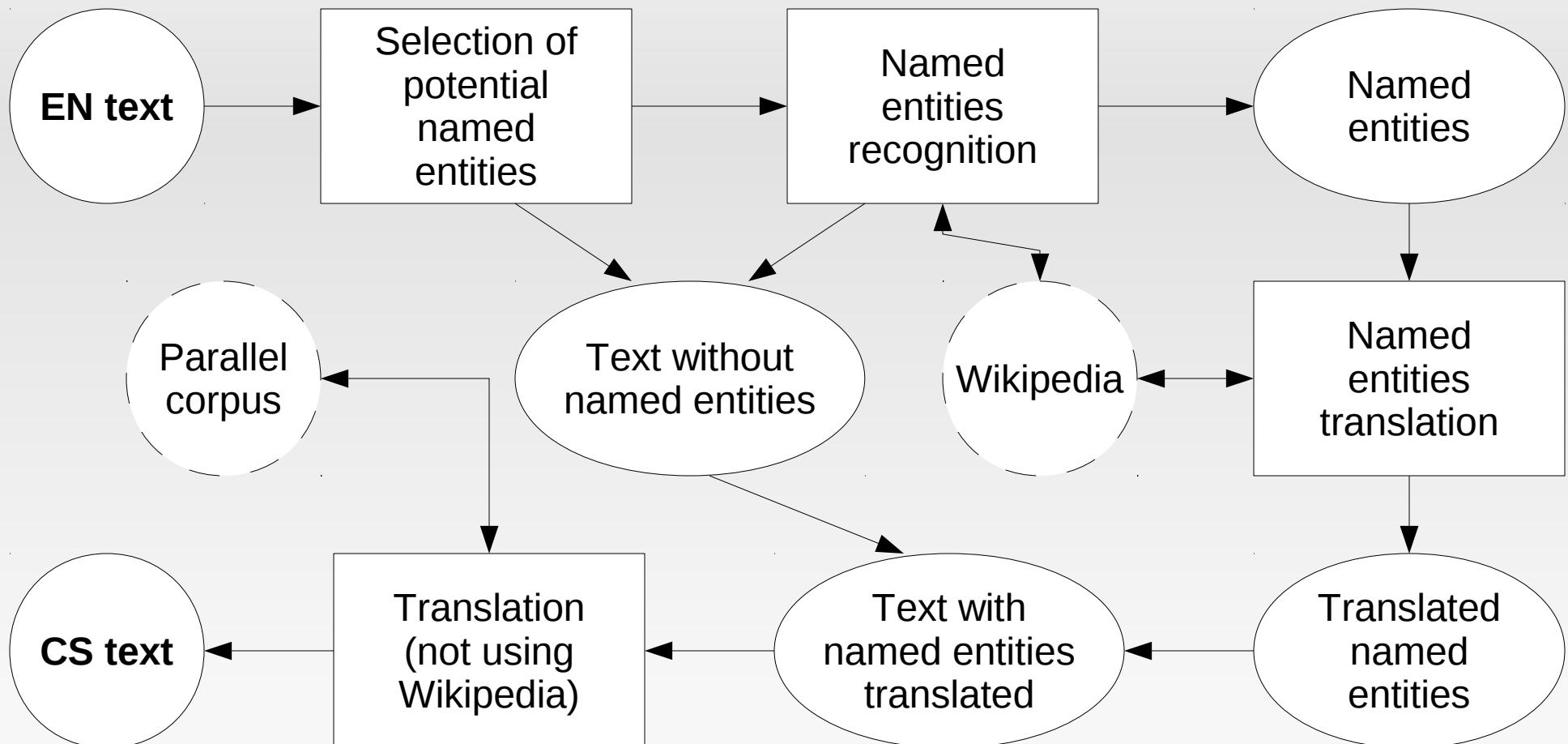
Překlad (česky > anglicky)

I live in London.

MT of NE with help of Wikipedia

- English to Czech translation
- Named entities recognition
 - look for possible named entities
 - filter the candidates using categories of the English article on Wikipedia
- Named entities translation
 - use title of the corresponding Czech article on Wikipedia
 - include inflected forms of the article name

Overview

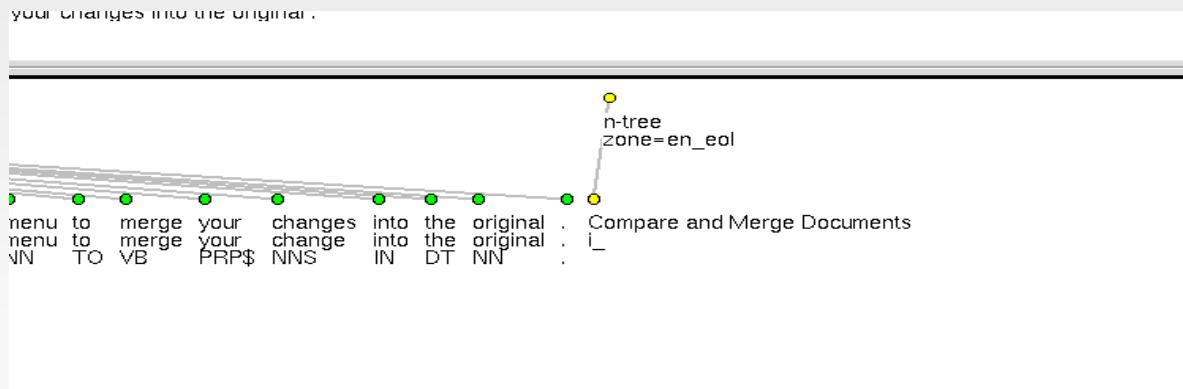


Named entity recognition

- Select phrases suspected to be named entities
 - “Rice University is at 6100 Main Street.”
- Human inter-annotator agreement only 83%
- Simple recognizer
 - Look for sequences of words with capital first letter
 - Use a small set of rules for beginnings of sentences
 - Precision 0.57, recall 0.73 against human annotation
- Stanford NER

Stanford NER

- More efficient recognition approach
- Capable of better multi-word identification
- TectoMT built-in feature
- Over 90% accuracy
- Precision 0.70, recall 0.49 against our human annotation



Named entity recognition

- Get categories of the article on Wikipedia
- Search superior categories (BFS)
 - Handmade list of categories containing named entities

Named entities categories

- Places
- People
- Organizations
- Companies
- Software
- Transport infrastructure

Get (all) categories



The Wikipedia logo, featuring a globe composed of puzzle pieces with various symbols and letters (Omega, W, N, etc.) on it.

WIKIPEDIA
The Free Encyclopedia

Article Discussion

The Call for Participation for Wikimania 2011 has been released. [Submit your presentation](#)

Rice University

From Wikipedia, the free encyclopedia

William Marsh Rice University, commonly referred to as **Rice University** or **Rice**, is a public research university located in [Houston](#), [Texas](#), United States. The university is located near the [Houston Museum District](#) and ad-

versity | [Educational institutions established in 1891](#) | Universities and colleges in the United States | North American Higher Education Collaboration | Universities and colleges

Category:Educational institutions established in 1891

Superior categories search

- **Educational institutions established in 1891**
- Educational institutions established in the 1890s
- Educational institutions established in the 19th century
- Educational institutions by year of establishment
- Organizations by year of establishment
- **Organizations**

Get categories – Wikimedia API

- `http://en.wikipedia.org/w/api.php?action=query&prop=categories&redirects&clshow=!hidden&format=xml&titles=Rice_University`
- `<?xml version="1.0"?>
<api><query><pages>
<page pageid="25813" ns="0"
 title="Rice University">
<categories>
 <cl ns="14" title="Category:Association
 of American Universities" />
 <cl ns="14" title="Category:Educational
 institutions established in 1891" />
 ...
</categories>
</page>
</pages>
</query>
</api>`

...

Named entity translation

Suppose that we have an English named entity:

- Look if there is the article on English Wikipedia
- Look if there is an equivalent Czech article
- Use all inflected occurrences of the translated name in the Czech article

Translation of “Spain”

IPEDIA
Encyclopedia

ge
s
content
events
article
to V

The Call for Participation for Wikimania 2011 has been releas

Spain **1**

From Wikipedia, the free encyclopedia

This article is about the country. For other uses, see Spain.

Spain (UK /'speɪn/ spayn; Spanish: **España**, pronounced [es'panya]) is a member state of the European Union and east by the Mediterranean Sea.

Cebuano

Česky **2**

Chamoru

Ch Španělsko

Zahiboanya

PEDIA
encyklopedie
na
pedie



Španělsko **3**

Španělsko, oficiálně Španělské království (španělsky Reino de España, castilian Spanish: Reino de España, Spanish: Reino de España) je stát ležící na Pyrenejském poloostrově, na severu s Francií a na jihu s Gibraltararem; španělské sev

Reading the articles

- Fetch the article content using WikiMedia API
- Ignore Wiki markup
- Trim last 3 letters of each word in the name, look for identical sequences of “stems” in the article
- Estimate the probability of different forms from the count of occurrences

Named entities in Moses

- Include our translation suggestions in the input data using XML markup
- Easy to incorporate alternative translations and their probabilities
- $p(e|f)$, $p(f|e)$, $\text{lex}(e|f)$, $\text{lex}(f|e)$ replaced by our score
 - only intended to differentiate between our suggestions
 - phrase table entries have much lower scores
 - LM sorts this out sometimes
 - needed to supply nonzero probability for OOV in LM

Experimental setup

- CzEng 0.9 corpus
- 200k parallel sentences for TM
 - alignment computed on 4-letter word stems
- 5m target-side sentences for LM
 - different from the parallel data
- All data lowercased
- Tools used: SRILM, GIZA++, Moses decoder & toolkit, eman

Experimental results

- Several combinations are possible with our approach:

Input preprocessing	BLEU
only title translation, force our translations, force untranslated unknown* forms	25.13
only title translation, force our translations	25.38
only title translation, allow phrase table translations	25.80
all name forms & probabilities, allow phrase table translations	25.97
same as above, Stanford NER	25.98
none	26.62

* *Named entity, but Czech article does not exist*

Experimental results

- Our method did sometimes improve translation:

Source: It was **Nova Scotia** on Wednesday.

Baseline: byl to **nova scotia** ve středu.

NE translated: to bylo **nové skotsko** ve středu.

Source: In August, 1860, they returned to the **Victoria Falls**.

Baseline: v srpnu, 1860, se k vyjádření **falls**.

NE translated: v srpnu, 1860, se na **viktoriiny vodopády**.

Sources of errors

- Wrong Wikipedia translation
 - the article is about a different meaning of the term
 - Brussels -> Bruselský region (Brussels Region)
- Failure of suffix trimming
 - the heuristics for searching in Wikipedia articles matches an unrelated term
 - Polsko -> pole (field), Nestlé -> nesprávně (incorrectly)
- Wrong named entity form
 - the article does not include the suitable inflected form
 - the language model fails to enforce the correct option

Manual evaluation

- 255 sentences, roughly 400 named entities
- No reference translations => BLEU not possible
- 3 systems:
 - baseline
 - translate unknown entities
 - keep unknown entities untranslated
- Outputs randomized by QuickJudge, ranked by 4 annotators, ties were allowed
- Translations differed only in 78 sentences

Manual evaluation

- All 4 annotators agree on a winner in about 25% cases
- Number of wins of each system:

Annotator	Baseline	Translate unknown	Keep unknown
1	46	56	51
2	38	45	54
3	41	39	47
4	35	43	49

Conclusion

- Improvement in some translations
- Drop in BLEU due to high number of errors
- Human evaluation favourable
- Future work:
 - better model probabilities of our translations
 - explore other ways of incorporating NE translations
 - improve NER

References

- Ondřej Bojar: *NPFL087 Statistical MT*
<http://www1.cuni.cz/~obo/vyuka/>
- Wikipedia, The Free Encyclopedia:
Named entity recognition
 - http://en.wikipedia.org/wiki/Named_entity_recognition
- MediaWiki: *MediaWiki API documentation*
 - http://www.mediawiki.org/wiki/API:Main_page
- Ondřej Bojar, Zdeněk Žabokrtský: *CzEng, Large Parallel Treebank with Rich Annotation*
 - <http://ufal.mff.cuni.cz/czeng/>