Source-context Features for English-to-Czech Machine Translation

Kamil Kos

41903676

Supervisor: Mark Dras

November 13, 2009

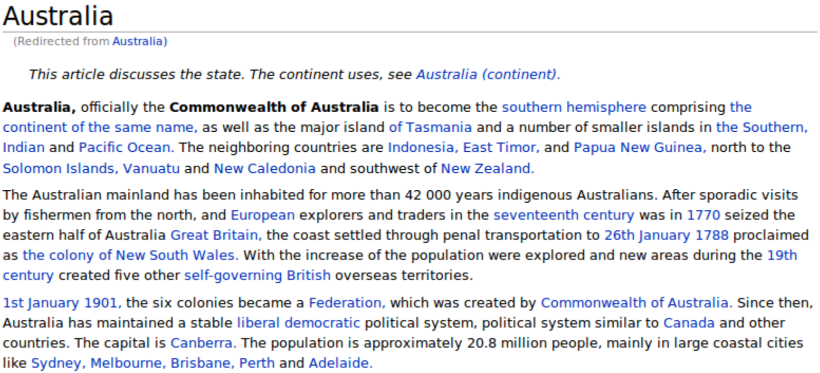
Outline

1. Introduction
2. Statistical MT
3. Past Work
4. Selected Features
5. Conclusion

Machine Translation (MT) - Why?

- difficult to learn a foreign language
- too many languages
- formal description of language
  - can computers understand language?
- cheaper than human translators

MT Example



Australia

(Redirected from Australia)

This article discusses the state. The continent uses, see Australia (continent).

Australia, officially the **Commonwealth of Australia** is to become the southern hemisphere comprising the continent of the same name, as well as the major island of Tasmania and a number of smaller islands in the Southern, Indian and Pacific Ocean. The neighboring countries are Indonesia, East Timor, and Papua New Guinea, north to the Solomon Islands, Vanuatu and New Caledonia and southwest of New Zealand.

The Australian mainland has been inhabited for more than 42 000 years indigenous Australians. After sporadic visits by fishermen from the north, and European explorers and traders in the seventeenth century was in 1770 seized the eastern half of Australia Great Britain, the coast settled through penal transportation to 26th January 1788 proclaimed as the colony of New South Wales. With the increase of the population were explored and new areas during the 19th century created five other self-governing British overseas territories.

1st January 1901, the six colonies became a Federation, which was created by Commonwealth of Australia. Since then, Australia has maintained a stable liberal democratic political system, political system similar to Canada and other countries. The capital is Canberra. The population is approximately 20.8 million people, mainly in large coastal cities like Sydney, Melbourne, Brisbane, Perth and Adelaide.

Source: `www.translate.google.com`

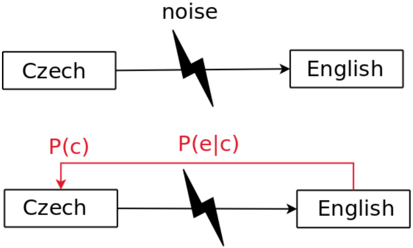
Source-context Features for English-to-Czech MT

- source context
- feature functions
- English-to-Czech
- machine translation

Outline

1. Introduction
2. Statistical MT
3. Past Work
4. Selected Features
5. Conclusion

Noisy-channel Model



$$\hat{c} = \underset{c}{argmax}\, P(c|e) \quad (1)$$

$$= \underset{c}{argmax}\, P(e|c) \times P(c) \quad (2)$$

- $P(c)$ - language model
- $P(e|c)$ - translation model

Log-linear Model

$$\hat{c} = \underset{c}{argmax} \sum_{n=1}^{N} \lambda_n h_n(c, e) \tag{3}$$

The model includes:

- feature functions $h_n(c, e)$
- feature weights λ_n
- optimum search for the best translation c

Phrase-based SMT

- translates small chunks of text instead of words
- *good evening* → *dobrý večer*
- *good* can be translated into Czech *dobrá*, *dobré*, *dobrou*, *dobrým*, *dobrému*, *dobrého*, . . .
- considers only local syntactical relations
- long dependencies are ignored - e.g. relative clauses in German

Outline

1. Introduction
2. Statistical MT
3. Past Work
4. Selected Features
5. Conclusion

Word Sense Disambiguation (WSD)

1. external WSD module selects best translation [CW05]
2. additional log-linear feature [CNC07]
3. augmented phrase-tables [CW07]

Log-linear Model Features

- additional feature functions with automatically optimized weights λ_n
- can be divided into ([GS08]):
  - lexical context features (collocation)
  - shallow syntactic features (part-of-speech)
  - syntactic features (parse tree)
  - positional features
- use of combinatorial categorial grammar (CCG) tags [BO07]

Outline

1. Introduction
2. Statistical MT
3. Past Work
4. Selected Features
5. Conclusion

Observations

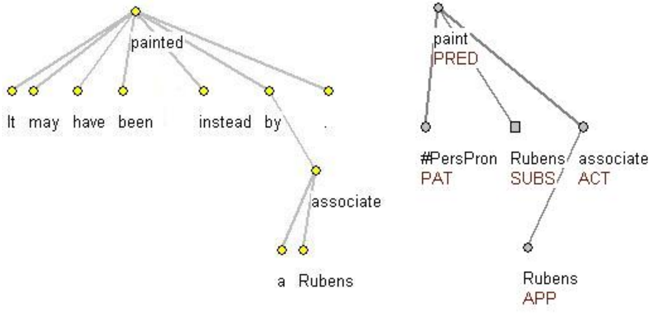
- external modules do not help much
- source context must be used directly in the MT model
- log-linear feature functions are the easiest way
- source context can improve MT quality across different languages

Features for English-to-Czech MT

- lexical collocations
- POS context
- syntactic features
- deep syntactic features

Deep Syntactic Features

- relations only between content words



It may have been painted instead by a Rubens associate.

Outline

1. Introduction
2. Statistical MT
3. Past Work
4. Selected Features
5. Conclusion

Project Achievements

- overview of existing approaches to source context in MT
- proposal of features suitable for English-to-Czech translation

Thank you for attention!

Questions?

- Alexandra Birch and Miles Osborne.
  CCG Supertags in Factored Statistical Machine Translation.
  In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16. Association for Computational Linguistics, June 2007.

- Yee Seng Chan, Hwee Tou Ng, and David Chiang.
  Word Sense Disambiguation Improves Statistical Machine Translation.
  In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*,

pages 33–40. Association for Computational Linguistics, June 2007.

Marine Carpuat and Dekai Wu.
Word Sense Disambiguation vs. Statistical Machine Translation.
In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 387–394, Ann Arbor, June 2005. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu.
Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation.
In *Proceedings of Machine Translation Summit XI*, pages 73–80, Copenhagen, Denmark, 2007.

Kevin Gimpel and Noah A. Smith.
Rich source-side context for statistical machine translation.
In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, Columbus, Ohio, June 2008. Association for Computational Linguistics.