METIS-II: a hybrid MT system

Peter Dirix

Vincent Vandeghinste
Ineke Schuurman
Centre for Computational Linguistics

Katholieke Universiteit Leuven

TMI 2007, Skövde

Overview

- Techniques and issues in MT
- The METIS-II project
- Intermediate evaluation and ongoing work

Overview of techniques in MT

- Since 50s: word-by-word systems
- Later: rule-based systems (RBMT)
- Since 80s: statistical MT (SMT)
- 90s: example-based MT (EBMT)

Issues

- SMT/EBMT need huge parallel corpora with aligned text (often not available)
- SMT/EBMT sparsity of data
- RBMT infinity of rules/vocabulary → manual work, nearly impossible
- RBMT advanced analytic resources needed

Resolve issues

- Use only large monolingual corpora (widely available)
- Use basic analytic resources and an electronic translation dictionary
- Enable construction of new language pairs more easily
- Combine EBMT/SMT and RBMT techniques to resolve disjoint issues
- Construct **hybrid MT system**

The METIS-II Project

- European project consisting of KULeuven, ILSP Athens, IAI Saarbrücken, and FUPF Barcelona
- Language pairs Dutch, Greek, German and Spanish to English
- Ongoing work (2004-2007)
- Build further on an assessment project (2002-2003)

Three language models

- Source-language model (SLM): analyses the structure in SL – tokenizers, lemmatizers, PoS taggers, chunkers, …
- Translation model (TM): models mapping between languages: dictionary, tag mapping rules, …
- Target-language model (TLM): uses TL corpus to pick most likely translation

Source-language model (Dutch)

- Tokenizer
- Tagger
- Lemmatizer
- Chunker

SLM: Tokenizer

- Rule-based tokenizer for Dutch
- 99.4% precision and recall

SLM: PoS tagger

- External tool: TnT (Brants 2000)
- About 96-97% accuracy for Dutch
- Trained on CGN (Corpus of Spoken Dutch)
- Uses CGN/DCoi tag set

SLM: Lemmatizer

- In-house, rule-based
- Uses tags and CGN lexicon as input
- Deals with separable verbs
- Future plans: use memory-based DCoi tagger/lemmatizer

SLM: Chunker

- In-house robust chunker/shallow parser: ShaRPa 2.1
- Steps can be defined as context-free grammars (non recursive) or perl subroutines
- Detects NPs, PPs and verb groups (F = 95%)
- Marks subclauses and relative clauses (F = 70%)
- Future plans: add subject detection

Translation model (Dutch to English)

- Bilingual dictionary
- Tag-mapping rules
- Expander (extra rules/statistics to deal with language-specific phenomena, e.g. reorganising word/chunk order, adding/deleting words,…)

TM: Dictionary

- Compiled from free internet resources and EuroWordNet
- About 38,000 entries and 115,000 translations
- XML format
- Contains relevant PoS and chunking information
- Contains complex and discontinuous entries

TM: Tag-mapping rules

- Mapping between Dutch (CGN/DCoi) and English (BNC) tag sets
- Uses mapping table

TM: Expander

- Generates extra translation candidates
- Deals with tense mapping
- Treats verb groups
- Inserts *do* when necessary
- Translates *like to* + infinitive
- Translates *om te* + infinitive

Target-language model (English)

- TL corpus preprocessing: same process as SL (tokenizing, lemmatizing, tagging, chunking,…) + draw statistics/put in DB
- TM has generated a list of possibilities
- Corpus look-up ranks possibilities according to TL corpus statistics
- Selects most likely translation or n-best
- Token generator for morphological generation

TLM: Corpus

- Corpus preprocessing: BNC (British National Corpus)
- BNC is already tokenized and tagged
- Lemmatized using IAI lemmatizer
- Chunked using ShaRPa 2.1 (NPs, PPs, VGs, subclauses, …)
- Put into SQL database

TLM: Corpus statistics

- Drawn statistics from corpus
- Co-occurrence of lemmas, chunks (heads), …
- Put into database

TLM: Corpus look-up (ranker)

- Dictionary look-up, tag-mapping rules, expander => result = bag of bags
- Lexical selection + word/chunk order is drawn from TL corpus
- Makes a ranking of candidate translations

Example (1)

- We want to translate: ‘De grote zwarte hond blaft naar de postbode’.

Example (2)

MATCHING WORDS	CORPUS INFO	FREQ
the/big/black/dog	the/big/,/black/lead/dog	1
the/large/black/dog	the/large/black/dog	1
the/big/dog	the/big/dog	20
	the/big/yellow/dog	4
	the/big/dog/party	1
	the/big/dog/'s/snarl	1
	…	
the/black/dog	the/black/,/tan/and/white/dog	1
	the/black/dog	20
	Churchill/and/the/black/dog	1
	…	
the/great/dog	the/great/dog	3
	…	
	…	
the/dog	*more than 1000 matches*	

Example (3)

SOLUTION	SCORE	freq	m	cumul(m)	NEW WEIGHT
the large black dog	1.000	1	4	2	0.707
the big black dog	0.667	1	4	2	0.472
the big gloomy dog	0.750	5	3	26	0.329
the grown up gloomy dog	0.500	18	2	76	0.243
the major gloomy dog	0.500	18	2	76	0.243
the great black dog	0.750	2	3	26	0.208
the tall black dog	0.750	1	3	26	0.147
the grown up black dog	0.750	1	3	26	0.147
the major black dog	0.750	1	3	26	0.147
the large gloomy dog	0.750	1	3	26	0.147
the black great dog	0.429	1	3	26	0.119
…					

Example (4)

BAG (HEADS)	RESULT	SCORE	freq	m
dog / bark / to / .	dog to bark .	0.267	2	4
	dog bark to .	0.222	1	4
	to bard dog .	0.190	1	4
dog / bark / at / .	dog bark at .	0.500	1	4
	dog at bark .	0.308	1	4
	at dog bark .	0.222	1	4
dog / bark / towards / .	towards dog bark .	0.267	1	4
	dog towards bark .	0.063	1	4
	dog bark towards .	0.286	1	4
dog / bark / toward / .	toward dog bark .	0.500	3	3
	toward bark dog .	0.143	1	3
	dog toward bark .	0.375	1	3
	dog bark toward .	0.600	1	3
	bark toward dog .	0.300	1	3
…				

Example (5)

SENTENCE	RESULT
the large black dog barks/bark at the postman .	0.00101608892330194
at the postman the large black dog barks/bark .	0.00101608892330194
the big black dog barks/bark at the postman .	0.00051978210288697
at the postman the big black dog barks/bark .	0.00051978210288697
the big gloomy dog barks/bark at the postman .	0.00037152767431080
at the postman the big gloomy dog barks/bark .	0.00037152767431080
the tall black dog barks/bark at the postman .	0.00028540695707770
at the postman the tall black dog barks/bark .	0.00028540695707770
the great black dog barks/bark at the postman .	0.00028243656500730
at the postman the great black dog barks/bark .	0.00028243656500730
the major gloomy dog barks/bark at the postman .	0.00022256538776012
at the postman the major gloomy dog barks/bark .	0.00022256538776012
the large black dog barks/bark to the postman .	0.00021386773758162
…	

Translation process

- Wrapper for whole process
- Analyse SL sentence(s)
- Build TM
- Pick translations with highest rank(s) and do token generation
- Offer translations to translator for post-editing (not implemented yet)

Evaluation

- Evaluated with BLEU, NIST and Levenshtein distance algorithm

	BLEU
average	0.3024
best	0.3486

Ongoing work & ideas

- Reimplementing the system (code clean-up)
- Elaborate rules (e.g. continuous tenses), lexica, …
- Take SL chunk order into account
- Improve SL and TL toolsets
- Provide tools for post-editing
- PACO-MT

Related work

- Context-based Machine Translation (CBMT, Carbonell 2006)
- Generation-heavy Hybrid Machine Translation (GHMT, Habash, 2003)

Questions

- ?