Applications
of Machine Translation

Index

- Historical Overview
- Commercial Products
- Open Source Software
- Special Applications
- Future Aspects

History

Before the Computer:

Mid 1930s: Georges Artsrouni and Petr Troyanskii applied for patents for ‘translating machines’

Proposing

- a method for an automatic bilingual dictionary
- a scheme for coding interlingual grammatical roles (based on Esperanto)

History

The pioneers, 1947-1954

first appearance of ‘electronic calculators’
research on using computers for translating natural languages

Motivations:

- wartime successes in code breaking
- developments by Claude Shannon in information theory

1954 the first public demonstration (IBM and Georgetown University)

History

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

History

The decade of optimism, 1954-1966

Early systems:

- bilingual dictionaries
- rules for producing correct word order

  (rules for syntactic ordering were too complex, need for more systematic methods of syntactic analysis)

"breakthroughs" , "semantic barriers" ,no solution

History

Automatic Language Processing Advisory Committee (ALPAC)

- famous 1966 report: MT slower, less accurate and twice as expensive as human translation
- no need for further investment in MT research

History

After the ALPAC report, 1966-1980

- virtual end to MT research in the United States, Soviet Union
- Research continues in Canada, France Germany

History

SYSTRAN System

- USAF (1970)
- Commission of the European Communities (1976)

METEO (Montreal University)

- Translation of weather reports

History

Motivation:

1960s:

English-Russian translation of scientific and technical documents

Mid-1970s:

administrative and commercial demands of multilingual communities and multinational trade (Europe, Canada, Japan)

History

The 1980s:

mainframe systems

- Logos (German-English and English-French)
- Pan American Health Organization (Spanish-English and English-Spanish)
- Metal system (German-English)
- major systems for English-Japanese

History

Microcomputers (text-processing) -> cheaper MT systems

Research -> advanced methods (indirect translation, interlingual)

Projects: GETA-Ariane (Grenoble), SUSY (Saarbrücken), Mu (Kyoto), DLT (Utrecht), Rosetta (Eindhoven), Carnegie-Mellon University (Pittsburgh), Eurotra (EC supported), Japanese CICC

History

Early 1990s:

- IBM Group: Candide (statistical methods)
- Japanese Groups (corpora, example/based)
- Start of speech translation (speech recognition, synthesis and translation)

Projects: ATR (Nara, Japan), JANUS (ATR, Carnegie-Mellon University, University of Karlsruhe), Verbmobil (Germany)

History

Changing focus:

“pure” research -> practical applications

- translator workstations for professional translators
- translating components in multilingual information systems

History

Late 1990s and early 2000s:

- software localisation
- growth in sales of MT software for PC
- MT from online networked services

  (e.g. AltaVista)
- automatic translation for direct Internet applications (electronic mail, Web pages, etc.), fast, less quality

SYSTRAN

- Founded in 1968
- 1970: US Air Force (Russian-English)
- 1974/75: NASA (joint Apollo-Sojus project)
- 1975: Prototype for European Commission
- 1978: Xerox (multilingual products)
- 1981: Japanese-English
- 1989: Customer Specific Dictionaries

SYSTRAN

- 1992: “C” conversion project (PC)
- 1996: US National Air Intelligence (Eastern European languages)
- 1997: embedded MT software (Ford), Babelfish
- 1998: Online Gaming (EA)
- 2000: OracleMobile (wireless potal)

- Yahoo!, Altavista Babelfish, Google translate

SYSTRAN

Daimler-Chrysler:

1998 merger of Daimler-Benz AG with the Chrysler Corporation>more than 372,500 people in 37 countries>Communication Challenge!

Professional Translation for Human Resource Materials, etc.

BUT:

SYSTRAN

MT-System for informal Communication (emails,…)

Features:

- German-English bidirectional language pairs
- No installation of client software
- Seamless integration with DC’s IT environment
- Low performance costs
- No maintenance requirements
- Ease of use and access

SYSTRAN

Implementation:

- Intranet Installation used by 25,000 employees for translation of Web pages, emails and corporate documents
- Browser-based interface
- Central SYSTRAN intranet server

SYSTRAN

Autodesk: Multilingual Customer Support

Implementation:

SYSTRAN hosts translation servers, develops the specialized lexicons, glossaries, and graphs for the highly technical Autodesk vocabulary

System is expected to produce usable quality with no postediting

SYSTRAN

Cinématique Gaumont: French film library

For universal access SYSTRAN provides a translation solution for researchers to access the database in English

SYSTRAN

Today:

XML-based engine

3 dictionaries:

- User dictionary
- Translation Memories (strength of human translation)
- Normalization dictionary

Pharaoh

- A beam search decoder for phrase-based models
  - works with various phrase-based models
  - beam search algorithm
  - time complexity roughly linear with input length
  - good quality takes about 1 second per sentence
- Very good performance in DARPA/NIST Evaluation
- Freely available for researchers http://www.isi.edu/licensed-sw/pharaoh/

Pharaoh

Components needed:

- Decoder code
- Parallel Corpus (e.g. Europarl: from the proceedings of the European Parliament)
- Training Program for language model
- Additional Tools (word lattices, n-best list,…)
- Training System to generate translation models

Pharaoh

Running the decoder:

```
% echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini > out
Pharaoh v1.2.9, written by Philipp Koehn
a beam search decoder for phrase-based statistical machine translation models
(c) 2002-2003 University of Southern California
(c) 2004 Massachusetts Institute of Technology
(c) 2005 University of Edinburgh, Scotland
loading language model from europarl.srilm
loading phrase translation table from phrase-table, stored 21, pruned 0, kept 21
loaded data structures in 2 seconds
reading input sentences
translating 1 sentences.translated 1 sentences in 0 seconds
[3mm] % cat out
this is a small house
```

Pharaoh

Phrase Translation Table

der ||| the ||| 0.3
das ||| the ||| 0.4
das ||| it ||| 0.1
das ||| this ||| 0.1
die ||| the ||| 0.3
ist ||| is ||| 1.0
ist ||| ’s ||| 1.0
das ist ||| it is ||| 0.2
das ist ||| this is ||| 0.8
es ist ||| it is ||| 0.8
es ist ||| this is ||| 0.2
ein ||| a ||| 1.0
ein ||| an ||| 1.0
klein ||| small ||| 0.8
klein ||| little ||| 0.8
kleines ||| small ||| 0.2
kleines ||| little ||| 0.2

Pharaoh

Trace

% echo ’das ist ein kleines haus’ | pharaoh -f pharaoh.ini -t

[...]

this is |0.014086|0|1| a |0.188447|2|2| small |0.000706353|3|3|

house |1.46468e-07|4|4|

- Trace for each applied phrase translation:
  - output phrase (there is)
  - cost incurred by this phrase (0.014086)
  - coverage of foreign words (0-1)

Pharaoh

Reordering

% echo ’ein kleines haus ist das’ | pharaoh -f pharaoh.ini -t -d 0.5

[...]

this |0.000632805|4|4| is |0.13853|3|3| a |0.0255035|0|0|

small |0.000706353|1|1| house |1.46468e-07|2|2|

- First output phrase this is translation of the 4th word

Pharaoh
Hypothesis Accounting

% echo ’das ist ein kleins haus’ | pharaoh -f pharaoh.ini -v 2
[...]
HYP: 114 added, 284 discarded below threshold, 0 pruned, 58 merged.
BEST: this is a small house -28.9234

- Statistics over how many hypothesis were generated
  - 114 hypotheses were added to hypothesis stacks
  - 284 hypotheses were discarded because they were too bad
  - 0 hypotheses were pruned, because a stack got too big
  - 58 hypotheses were merged due to recombination
- Probability of the best translation: exp(-28.9234)

Pharaoh

Many more Parameters to customize:

- Translation Options
- Future Cost Estimation
- Hypothesis Expansion
- Beam Size (trade-off between speed and quality)
- Limits on reordering
- Word Lattice Generation
- N-best list

Moses

Very similar to Pharaoh

Advanced Features:

- Lexicalized Reordering Models
- Binary Phrase Tables with On-demand Loading
- Efficient Language Model Handling
- Conversion to binary format
- Binary language model format
- Quantized language model format
- XML Markup
- Parameters

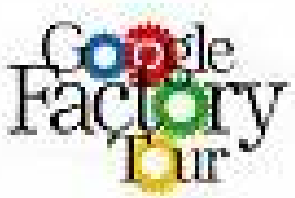
Developments

- Google Machine Translation Systems (Google Research Lab)

underlying principle:

system is learning from existing human translations , large corpus of texts (Rosetta Stone approach)

Developments



Original (Arabic)

البيت الابيض يؤكد وجود شريط مسجل جديد لبن لا

Existing translation

Alpine white new presence tape registered for coffee confirms Laden

Google Research translation

The White House Confirmed the Existence of a New Bin Laden Tape

Advanced Applications

AppTek

Speech-to-Speech Translation

- SpeechTrans™:
  integrated Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) with Machine Translation

  real-time dynamic speech-to-speech machine translation (computers, wearable machines, telephony servers)

Advanced Applications

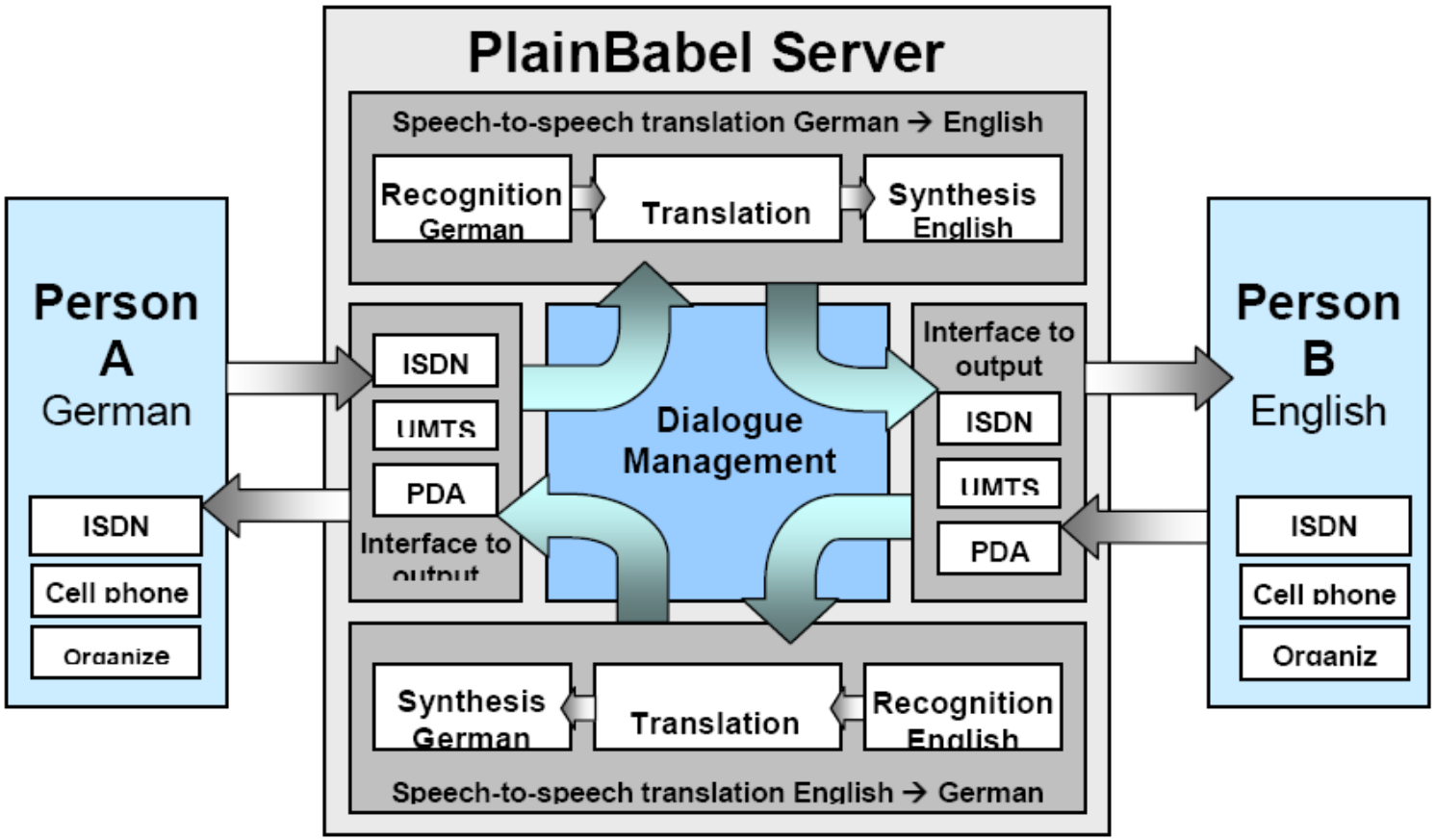
Apptek

PlainBabel – Simultanseous Machine Translation

dialogue driven speech-to-speech translation system that allows for a bilingual natural dialogue between two parties that don’t speak each others language. The system can be used with different types of output devices from an ISDN telephone to a 3G mobile phone.

Advanced Applications

Apptek



Advanced Applications

Apptek

Information Retrieval

- TextFinder:

  Linguistic enhanced information retrieval with query translation, thematic search, and name/event retrieval. Included morphology enhanced text indexing and searching.

Advanced Applications

Apptek

Media-mining

- Media Sphere “Telephony/TV” :

  Real-Time telephony and video broadcast capture, transcription, indexing, searching and playback. Includes Machine Translation engine for translation of transcribed text.

Advanced Applications

Apptek

Automatic Video Subtitling

Example: original German broadcast with English subtitles that were created by first running Speech Recognition to produce the German transcript and then Machine Translation to create English text.

Future

Trends:

- Highly integrated translation systems
- hybrid statistical-linguistic translation
- multi-engine translation systems
- development of new statistical techniques
- Portable speech-to-speech MT systems

LINKS

http://ourworld.compuserve.com/homepages/WJHutchins/
http://www.systransoft.com/
http://www.isi.edu/publications/licensed-sw/pharaoh/
http://www.statmt.org/moses/
http://www.apptek.com/
http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/tutorial2006.pdf
http://www.isi.edu/natural-language/publications.html
http://www.lsi.upc.edu/~jgimenez/MT/
http://www.geocities.com/langtecheval/
http://www.ikp.uni-bonn.de/dt/lehre/materialien/maschueb/
http://www.amtaweb.org/
http://www.eamt.org/
http://www.aamt.info/