Historical Survey of Machine Translation in Eastern and Central Europe

John Hutchins

Hamburg, May 2012

Outline

Pre-1990

- USSR, predominantly Russia
- Rule-based: Russian as target

Post-1990

- Variety of countries, no Russian dominance
- Corpus- and statistics-based
- Translation tools and resources

The beginnings

News of the Georgetown-IBM demonstration, January 1954

Petr Troyanskii's patent 1933 – [reported by Zhirkov in 1956]

Three groups formed in 1955/56

Yurij Panov at ITMVT

Alexei Lyabunov at Steklov Mathematical Institute

Nikolai Andreev at Leningrad University

Troyanskii

- Patent in 1933
- Pre-computer mechanical/electronic device
- Sloping table with dictionary of words in 6 languages
- Endings and prepositions replaced by Esperanto-type codes
- Discussed at USSR Academy of Sciences 1944, and rejected

ITMVT

[Institut Tochnoi Mekhaniki I Vychislitel'noj Tekhniki – Institute of Precise Mechanics and Computing Technology]

KGB backed. Director: Dmitrij **Panov** (author: first MT book in Russian, 1956

- Chief linguist: Izabella K. **Bel'skaya**

English-Russian

Direct translation – initially modelled on GU-IBM Russian-English system – ad hoc rules

Applied mathematics and (!) literature (Dickens, Edgar Allen Poe, etc.)

ITMVT – ad hoc rules

Interpretation of ***much***

1(2,3) check immediately preceding word for ***how***

2(0) skol'ko (numeral, invariable)

3(4,5) check immediately preceding word for ***as***

4(0) *stol'ko zhe (numeral, variable)*

5(7,9) *c*heck given word for ***much***

6(0) not to be translated (adverb)

7(6,11) check immediately preceding word for ***very***

8(0) *mnogii* (adjective, hard stem, with sibilant)

etc.

AMPAR

- KGB project (Yurij **Motorin**, Yurij **Marchuk**) – with students from Moscow Lomonosov State University
- Direct translation, English-Russian
- 17 stages including: dictionary lookup, morphological analysis, idioms, grammatical analysis, syntactic analysis, translation of unambiguous words, translation of ambiguous words (using contextological dictionary), grammatical analysis, control of intermediary text, morphological synthesis, text output
- Contextological dictionary (from corpora concordances)

Steklov Mathematical Institute

- Lyapunov and cybernetics
- Olga **Kulagina**
- French-Russian
- Multiple passes for morphological and syntactic information: verbs, nouns, pronouns, etc.
- Multi-word, collocations, idioms
- 17 elementary operators (basic algorithm)
- Set theory

Steklov Mathematical Institute

Tat'yana **Moloshnaya**

English to Russian

Matching by phrases/collocations

Different approach – more emphasis on syntax (Fries, Jespersen)

Adopted widely within and outside USSR

In 1967 Kulagina began new French-Russian system (FR-II)

Dependency, transfer-based

Institute of Linguistics

Igor **Mel'chuk**

Hungarian to Russian

Initially followed French-Russian model, but concluded deeper level of analysis needed

Origins of his meaning-text model

Interlingua: language-independent 'lexical functions' (verb and agent (*write*, *writer*), noun and inceptive verb (*war*, *break out*), noun and causative (*foundations*, *lay*)

Explanatory-combinatorial dictionary

Institute of Foreign Languages

Founded by Viktor **Rozentsvejg**

Yurij **Apresjan**

- 1968: sent to Institute of Heavy Electrical Machinery, Institute for Information Transmission Problems
- ETAP – based on meaning-text model (**Boguslavskij**)

Leningrad State University

Nikolaj **Andreev**

Multiple bi-directional systems: Rumanian, German, Norwegian, Serbocroat, Czech, Hindi, Indonesian, Turkish, Chinese, Arabic, Indonesian, etc.

Interlingua (independent language) based on most frequent lexical and syntactic features, weighted according to importance of source language (major languages more than minor)

Founded 1958; ended 1961

Speech Statistics Group

Founded 1962

Practical MT – statistical foundations

Pragmatic – recognition of limitations of computational approaches to natural language

NL (fuzzy, open, dynamic, polysemantic), computers (discrete, static, deterministic, rigidity

Raimund **Piotrovskij** (Leningrad)

Groups (branches): Kazakhstan, Kiev, Kishinev, Samarkand, etc.

All-Union Translation Centre

Set up 1974

First director: **Marchuk**; later: **Oubine**

Incorporated: AMPAR (English-Russian direct translation)

NERPA (German-Russian direct translation) from Inst Foreign Languages (**Martem'janov**)

- Later merged with AMPAR as ANRAP

FR-II (French-Russian transfer) from **Kulagina**

SILOD from Leningrad State U

ETAP-II (**Apres'jan**)

Operational systems from former USSR

STYLUS (later PROMT), founded by Svetlana **Sokolova** (SSG Leningrad)

Initially English-Russian, Russian-English, German-Russian

PARS founded by Mikhael **Blekhman** (SSG Kharkov)

Initially Russian-Ukrainian, English-Ukrainian

Central Patent Office (English-Russian) since 1964

MT research outside Russia in Soviet era

1958 conference in Moscow: 340 participants from 79 institutions

Very wide range of languages, but most came to nothing; individuals with no financial backing

But:

Some MT work on Georgian, Armenian, Ukraine (help from Mel'chuk and Moloshnaja)

German Democratic Republic

Akademie der Wissenscahften (**Agricola**, **Kunze**):

English-German, Russian-German (after mid 1980s working system abandoned); plans to be attached to METAL group

Czechoslovakia/Czech Republic

Charles University

Pre-1990: Petr **Sgall** (and Hajičova)

- Functional-generative, stratificational, dependency
- Zdeněk **Kirschner**
- APAC (English-Czech), based on Montreal Q-system

post-1990: Hajič, Bojar, Homola, Kuboň, etc:

- Dependency-based SMT, closely related languages

Hungary

pre-1990

Nothing after Mel'chuk except theoretical work by Ferenc **Papp** and Ferenc **Kiefer**

Post-1990

Morphologic (Gabor **Proszéky**)

Bulgaria

Pre-1990:

Alexander **Ljudskanov**: theoretician, widely known

Post-1990

Knowledge-based translator workstation

(Galia Angelova, Walter **von Hahn**)

Romania

Post-1990

Menu-driven translation aid (Cristina **Vertan**,

Walter **von Hahn**)

Institute for Artificial Intelligence

Alignment, disambiguation, SMT (Dan **Tufis**)

Poland, Slovenia

Poland

Krzysztof **Jassem**: Polish-E transfer (POLENG)

Slovenia

Language resources (MULTEXT-EAST, JRC-Acquis):

Tomasz **Erjavec**

Latvia, Lithuania, Estonia

Latvia

- Resources for under-resourced languages
- Morphology of Baltic languages
- Cloud-based platform (LetsMT!)

Lithuania

- Language resources

Estonia

- Parallel corpora, SMT

References

Paper on my website:

www.hutchinsweb.me.uk

Resources:

www.mt-archive.info