ARIANE (GETA) MT System

Presenter: Batuhan Baykara
Outline

1 Historical Background

2 The System

3 Processes and Components
   3.1 Application Environment
   3.2 Analysis Process
      3.2.1 Morphological Analysis
      3.2.2 Structural (Multi-Level) Analysis
   3.3 Transfer Process
      3.3.1 Lexical Transfer
      3.3.2 Structural Transfer
   3.4 Generation Process
      3.4.1 Syntactic Generation
      3.4.2 Morphological Generation
   3.5 Rule Writing Formalisms
      3.5.1 ATEF
      3.5.2 ROBRA

4 Tools Integrated to the System

5 Example Translations
Historical Background

- One of most fundamental MT systems
- GETA (Groupe d'Etudes pour la Traduction Automatique) former CETA (Centre d’Etudes pour la Traduction Automatique)
- Led by Bernard Vauquois
- Interlingua systems developed in CETA in 1960s (Russian-French)
- Renamed as GETA and Ariane system (transfer based) developed
- First release in 1978 (Ariane-78)
- Then other systems followed; Ariane-85, Ariane-G5
The Overall System

- Main goal: Create a workbench for linguists
- Transfer based system composed of 3 phases;
  - Analysis
  - Transfer
  - Generation
- Very complex system
- Used mostly in Russian-French translation
- But some German-French translations were made
- Other researchers that worked at some point in GETA experimented with English-Malay and English-Thai translations.
Application Process

- Before giving the input to the system, **pre-editing** can be done;
  - It is optional
  - Some problems are solved;
    - Mostly lexical ambiguities are solved
    - The antecedent of a relative pronoun
- After the translation is obtained, **post-editing** is possible;
  - This process can increase the quality of the translation significantly
  - It is an expensive process
  - Some sub-environment tools such as THAM are used
- Ariane is a **non-interactive** tool however; in some parts human interruptions may be necessary.
  - Correct spelling errors or make modifications to the dictionary
Analysis Process

- Two steps; morphological analysis and structural analysis.

Morphological Analysis

- Process the input according to ATEF formalism
- In the end a flat tree is produced.
- UL= Lexical unit, ULTXT=text, ULFRA= sentence, ULOCC=word
- Last level contains grammatical information..
Structural Analysis (Multi-level Analysis)

● Most complex and difficult part of the whole translation process
  ○ In depth analysis is required to find morphological, lexical and logico-semantic information.
    ■ Morphological level -> dogs (LU=dog and plural noun)
    ■ Syntactic level -> finding noun phrases, verb phrases etc..
    ■ Logico-semantic level -> deep syntactic representation showing dependency relations with their semantic roles (goal, cause, location, gender etc...)
  ○ The tree should be unambiguous at the end of the analysis process.
Analysis Process cont...

- In syntactic analysis ROBRA rule writing formalism is used.
- ROBRA is a tree-transducer system in the heart of Ariane.
- The system works as follows;
  1) Transformational rules (TR) are written by linguists
  2) These rules are grouped in transformational grammar (TG)
  3) TG is applied to the tree obtained from morphological analysis. Hence all TRs are executed on it
  4) The overall structure is control via a control graph which channels the input to the corresponding TGs.
- Additionally, other problems such as anaphora resolution besides ambiguity can also be resolved depending on the system configurations.
Transfer Process

Transfer phase consists of two steps; lexical transfer and structural transfer.

Lexical Transfer

TRANSFER component is used which is a bilingual multichoice dictionary of transfer rules. Takes the tree as an input and changes the labellings on the tree according to rules; it is like a pattern matcher.

- **Simple-to-simple substitution**: Directly translates the source lexical unit to the corresponding target lexical unit (one-to-one translation).
- **Simple-to-complex substitution**: A single source unit is translated into several target lexical units. For example, “avec” is translated as “by means of”.
- **Complex-to-simple or complex-to-complex substitution**: Multiple lexical units are translated as a single unit or multiple units.
Transfer Process cont...

Structural Transfer

- ROBRA is used at this step.
- Reconstruction of the source tree to the target tree structure is handled at this step.
- In this step, necessary alterations such as inserting or deleting is done.
Generation Process

The process consists of two steps; syntactic generation and morphological generation.

**Syntactic Generation**

- Takes the output obtained from transfer phase
- Computes the final surface syntactic structure
- Includes selection of appropriate verbal auxiliaries, rearrangement of word order and setting values of morphological variable values such as sumber and gender agreements.
- Again ROBRA is used
Generation Process cont...

**Morphological Generation**

- This is the last step of the translation system
- The output text is generated from the surface representation
- SYGMOR module is used
  - It is a rule writing formalism
  - Its function is to convert labelled tree structure into string format including punctuations.
  - It can be thought of as a decoder.
Rule Writing Formalisms

Ariane uses four different software packages which assist in the development of various phases.

- **ATEF**: String to tree transformation package used in **Analysis phase**
- **ROBRA**: Tree to tree transformation package used in **all phases**
- **TRANSF**: Tree to tree transformation package used in **Transfer phase**
- **SYGMOR**: Tree to string transformation module used in **Generation phase**
ATEF

- ATEF aims to handle the mappings of strings and convert them into a bunch of feature that are represented in as structured tree format.

```
LU=CONDENSE, CAT=V, CONJ=NO, TNS=PAST, VOX=PAS
$X + "D" ==
    [LU=$X, CAT=V, CONJ=NO, TNS=PAST, VOX=PAS] /
CAT($X)=V and end($X)="E"
```

- $X$ is a variable that ends with e and is in the lexicon category of verb (V).
- ATEF uses dictionary lookups to find the morphs.
Rule Writing Formalisms

ROBRA

- An example Transformational rule (TR) for compound nouns.
- -E- => equal
- -ET- => and
- -OU- => or
- -SI- => if
- -NE- => not equal
- -ALORS- => then
- -SINON- => else
- -FSI- => end if

```plaintext
COMPN: 1(2(3,4)) /

CAT(3) -E- N -ET-
(CAT(4) -E- N -OU- SUBD(4) -E- CARD)
-OU-
CAT(3) -E- PREF -ET- CAT(4) -E- N
==

1(3,4) /*<-->-2/
4:4,
-SI- SUBD(4) -NE- CARD
-ALORS- SF(4) := GOV,
SF(3) := JUXT,
RS(3) := QUAL,
VAR(1) := VAR(4)
-SINON- SF(3) := GOV,
SF(4) := JUXT,
RS(4) := QUAL,
VAR(1) := VAR(3)
-FSI-
1:1, K:NP, UL:='*NP', VLI:=N
```
Tools integrated to the System

- Ariane is designed to be a product therefore it needs to be working on all kinds of MT tasks.
- Some end-users, linguists, have requested extensions to the system.

**ATLAS:** A helper tool where linguist can add new words and rules to the dictionaries.

**THAM:** It is a text editor that can assist the linguist in the process of translation. It provides a dictionary which can be directly accessed from the screen. Importantly it provides a set of functions that are programmed which help linguist in terms of efficiency.
Tools integrated to the System

**VISULEX**: It is an easy to use visualization tool for assembling and separating essential information in the linguistic database. For instance, lexical database of Ariane is kept in more than 50 files, it is scattered all around. Hence, visulex makes it easier to access and see the lexical units.
Example Translations

- Ariane is mostly used in Russian-French,
- Tested on real world text
- Dictionaries used contains 7500 lexical units
  - 5000 in French
  - 2500 Russian
- The translations were made on an IBM mainframe.
- A total of 835 abstract and text were translated.
- Results were presented to the Ministry of Defence
LANGUES DE TRAITEMENT: RUB - FRB

TRADUCTION DU 6 NOVEMBRE 1986  8H 40MN 41S

VERSIONS : ( A : 21/07/86 - T : 21/07/86 - G : 21/07/86 )

-TEXTE SOURCE-
Cifrovaya obrabotka signalov v optike i gologralii. Vvedenie v cifrovuyu optiku.

Izlagayutsya osnovyi naukhnogo napraveliya, izukhayuthego ispolqzovanie cifrovykh processorov v optikheskix i golografikheskix sistemax. Rassmatrivayutsya voprosyi optimaiwnogo cifrovoego predstavleniya i modelirovaniya optikheskix signalov i fx preobrazovaniy, yeffrktivnye vykhislitelnye algoritmy i adaptivnye metodyi obrabotki izobrazhenii, gologramm i interferogramm, sinteza gologramm i yelementov optikheskix sistem.

-TEXTE TRADUIT-
Traitement numéral des signaux dans l'optique et la graphie nue. Introduction dans une optique numérique.

On expose les bases de la direction scientifique qui étudie l'utilisation de processeurs numéraux dans des systèmes optiques et nu (Genre-Nombre?) graphiques. On examine les problèmes de la représentation numérique optimale et du modelage de signaux opaques et de leurs transformations, algorithmes de calculateur efficaces et méthodes adaptables du traitement des représentations, des grammares nus et des interféromètres, de la synthèse des grammares nus et des...
SIMPOZIUM POSVYATEN NADERNOJ SPECTROSKOPII I STRUKTURE ATOMNOGO YADRA. V O VSTUPITELNOH SLOVE PODEKHIAETSYA VAZHNNAYA ROL' KOTOROIY SIMPOZIUM SVYGLAL V RAZVITII NADERNOY FIZIKI SLADVIX YENERGIY V SOVETSKOM SOYUZE. V XODE SIMPOZIUMA OBSUZHDENNY RYAD VAZHNYX ISSLEDOVANII OCHISTWLENNYX SOVETSKIMI UCHENNIKAMI. V KHASTNOSTI IZUKHEHO NESHRANENIY KHASTNOSTI V VAZHNYX PROCEHHAX SOZDAHNIY MODELI NEAKTYVNOGO YADRA SPOMNITAE DELJENIE IZOTOPOV SVERKITYAZHNYX YELEMENTOV I OBNARUZHENIE YEFKTE TENEJ PRI RASSEYANIY KHASTNY SOBRANNY YEDITELENNYIE STATISTIKHEKII DANNYYE, OTRAZHENNMMY ROST KHISSA PREDOLOZHENNYX DOKLADOV. OTHOMKHIAETSYA PRISETSTVIE SREDI UKHASTNYKH SPECIALISTOV IZ ZARUBLEYHIX STRAN.

--- ( TRADUCTION DU:-- MARS 1980 )

VERSIONS: ( A: 29/01/80; T: 12/09/79; G: 21/09/79 )

LE SYMPOSIUM L'EST CONSACRE A LA SPECTROSCOPIE NUCLEAIRE ET A LA STRUCUTRE DU NUCLEAS. DANS LE MOT D'ENTREE ON SOULIGNE LE ROLE IMPORTANT QUE LE SYMPOSIUM A JOUE DANS LE DEVELOPPEMENT DE LA PHYSIQUE NUCLEAIRE DES FAIBLES ENERGIES EN UNION SOVIETIQUE. PENDANT LE SYMPOSIUM ON A EXAMINE LA SERIE DES ETUDES IMPORTANTES REALISEES PAR LES SAVANTS SOVIETIQUES. EN PARTICULIER, ON A ETUDIE LA NON-CONSERVATION DE LA PARITE DANS LES PROCESSUS? PROCEDES NUCLEAIRES. DIVISION SPONTANEE DES ISOTOYPES DES ELEMENTS SUPERLORDS ET DECOUVERTE DE L'EFFECT DES OMBRES PENDANT LA DISPERSION DES PARTICULES. ON A REUNI LES DONNEES STATISTIQUES CONVAINCANTES QUI REFLETTENT LA CROISSANCE DU NOMBRE DES RAPPORTS PROPOSES. ON REMARQUE LA PRESENCE PARMi LES PARTICIPANTS DES SPECIALISTES DES PAYS ETRANGERS.
Thank you for listening...