Approaches to Machine Translation: Rule-based, Statistical and Hybrid

Rule-based Machine Translation

History (I)

- First RBMT systems developed in the 1970s:
  - Systran http://www.systran.de/
  - Japanese MT systems
    - » Toshiba Solutions http://hon-yaku.toshiba-sol.co.jp/
    - » NEC http://www.nec.co.jp/middle/meshplus/
    - » Fujitsu http://software.fujitsu.com/jp/atlas/
    - » LogoVista http://www.logovista.co.jp/
    - » IBM http://www-06.ibm.com/jp/software/internet/king/
    - » etc.

Source: http://en.wikipedia.org/wiki/Rule-based_machine_translation

History (II)

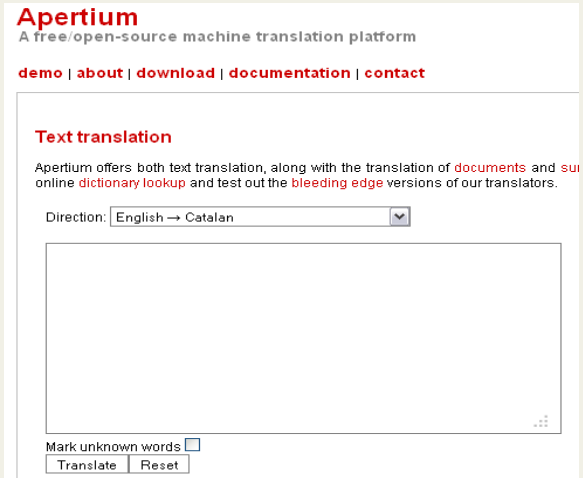
- EUROTRA (1978-1992, European Commission)
  - http://www-sk.let.uu.nl/stt/eurotra.html
  - MT for the 7-9 official languages

Source: http://en.wikipedia.org/wiki/Rule-based_machine_translation

Current Popular System

Apertium

http://www.apertium.org/

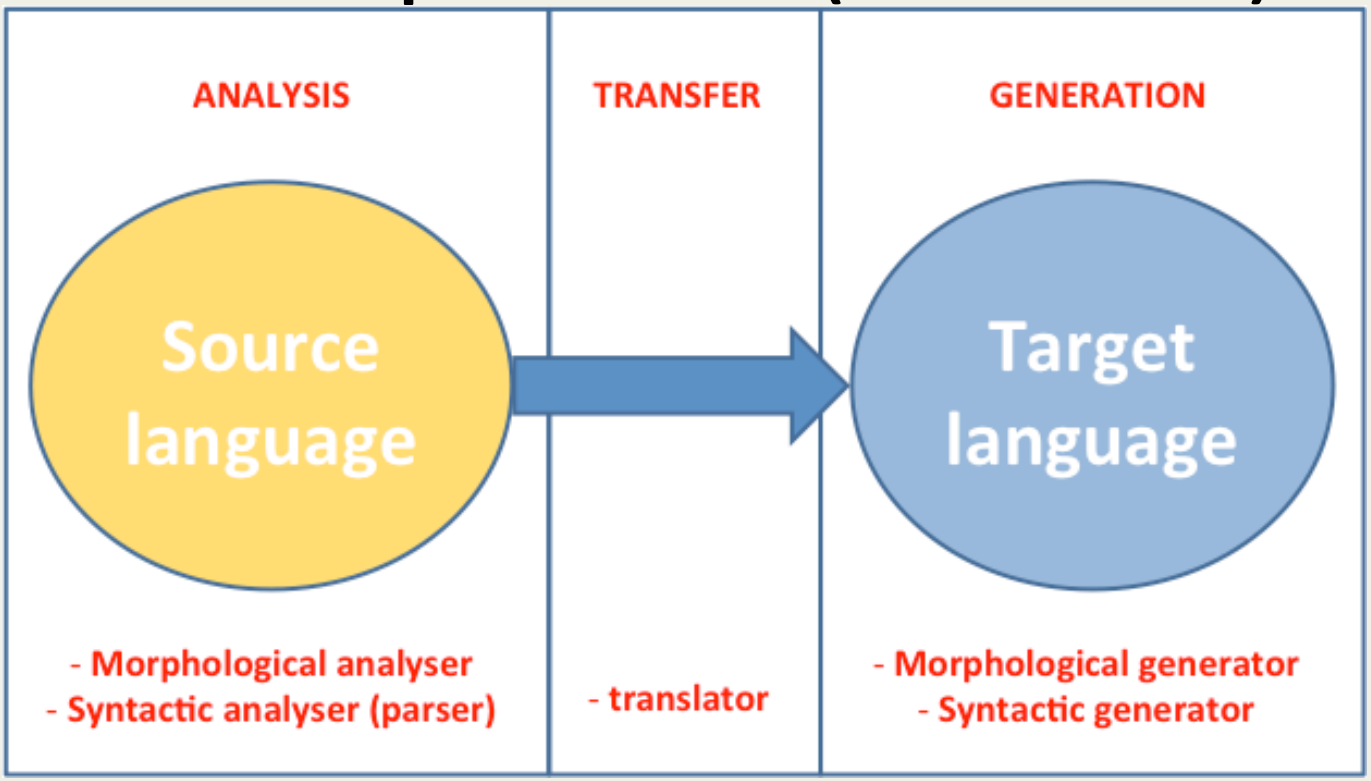


Licences and sources

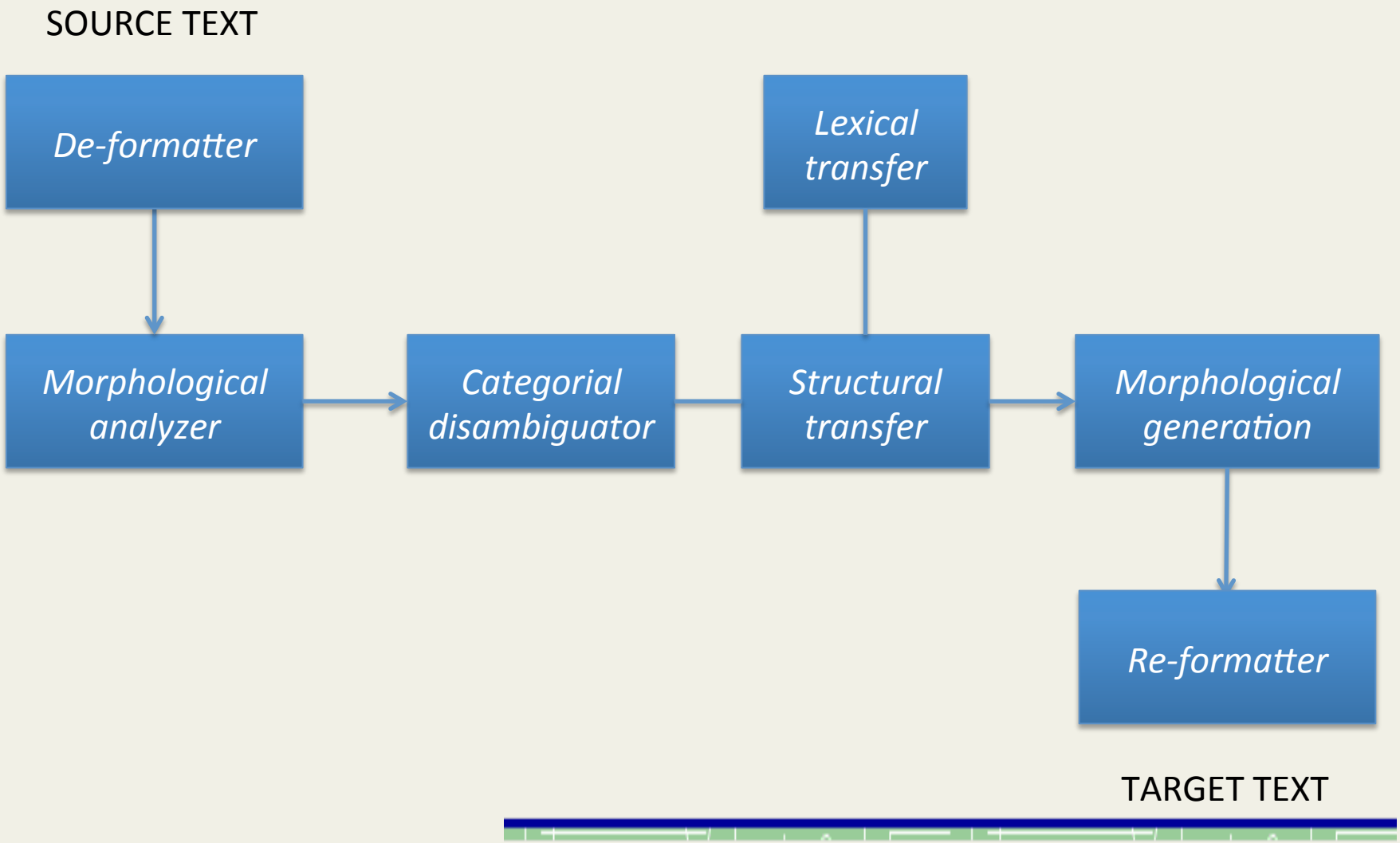
	License	Source available	
APERTIUM	**GPL**	**YES**	All programs and language data free and open source
SYSTRAN	**Commercial**	**NO**	Hybrid: rule-based & SMT

Source: http://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications

Components (Transfer)



Apertium: Block diagram



De-formatter

- Separates text from format information
- Based on finite-state techniques
- Available for: plain text, html, rtf, openoffice, word, ppt, excel, MediaWiki, QuarkXPress

$ echo "this is a test" | apertium-deshtml

[]this is[<\/em>]a[ ]test.[][<\/b>]

Morphological analyser

- Segments the source text in surface forms (SFs)
- Assigns to each SF one or more lexical forms (LFs): lemma, lexical category, morphological inflection information
- Processes contractions

Categorial disambiguator

- Picks one of the LFs corresponding to each ambiguous SF according to context
- Uses hidden Markov models and hand-written constraint rules
- Is trained using representative corpora for the source language

Structural transfer

- Rules have a pattern-action form
- It detects LF patterns to be processed using a left-to-right longest-match strategy
- It executes the actions associated to each pattern in the rule file to generate the corresponding LF pattern for the TL

More complex structural transfer

- For language pair with longer reorderings:
  - Patterns of LFs (chunks) are detected, processed and marked
  - Patterns of chunks are detected and processed: this interchunk processing allows for longer-range reorderings

Lexical transfer module

- Reads each SL LF and generates the corresponding TL LF

Morphological generator

- Generates from each TL LF a TL SF after adequately inflecting it

Post-generator

- Performs some TL orthographical transformations such as contractions
  - Can+not =cannot

Re-formatter

- Integrates format information into the translated text

Example of dictionaries

Monolingual (Spanish)

```
<e lm="cósmico">
  <i>cósmic</i>
  <par n="absolut/o__adj"/>
</e>
```

Monolingual (Catalan)

```
<e lm="còsmic">
   <i>còsmi</i>
   <par n="acadèmi/c__adj"/>
</e>
```

Bilingual dictionary (Spanish → Catalan)

```
<e>
   <p>
     <l>cósmico<s n="adj"/></l>
     <r>còsmic<s n="adj"/></r>
   </p>
 </e>
```

Example of transfer rule

This rule reorders
adjective + noun
Into
noun + adjective

In Chinese-to-Spanish

```
< rule comment = ”RU LE : adj nom” >
< pattern >
< pattern – itemn = ”adj”/ >
< pattern – itemn = ”nom”/ >
< /pattern >
< action >
< call – macron = ”f – concord2” >
< with – parampos = ”2”/ >
< with – parampos = ”1”/ >
< /call – macro >
< out >
< chunkname = ”j n”case = ”caseF irstW ord” >
        < tags >
                < tag >< lit – tagv = ”SN ”/ >< /tag >
                < tag >< clip pos = ”2”side = ”tl”part = ”gen”/ >< /tag >
                < tag >< clip pos = ”2”side = ”tl”part = ”nbr”/ >< /tag >
                < tag >< lit – tagv = ”p3”/ >< /tag >
        < /tags >
< lu >
        < clip pos = ”2”side = ”tl”part = ”whole”/ >
< /lu >
< b pos = ”1”/ >
        < lu >
                < clip pos = ”1”side = ”tl”part = ”lem”/ >
                < clip pos = ”1”side = ”tl”part = ”a adj”/ >
                < clip pos = ”1”side = ”tl”part = ”gen”/ >
                < clip pos = ”1”side = ”tl”part = ”nbr”/ >
< /lu >
< /chunk >
< /out >
< /action >
< /rule>
```

Question

(1) Given the following rule

```
<rule>
<pattern>
<pattern-item n="determinant"/>
<pattern-item n="adjectius"/>
<pattern-item n="nom"/>
</pattern>
<action>
<out>
<lu>
<clip pos="1" side="tl" part="lem"/>
<clip pos="1" side="tl" part="a det"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
<b/>
<lu>
<clip pos="3" side="tl" part="lem"/>
<clip pos="3" side="tl" part="a nom"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
<b/>
<lu>
<clip pos="2" side="tl" part="lem"/>
<clip pos="2" side="tl" part="a adj"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
</out>
</action>
</rule>
```

(2) Choose one interpretation

- Agreement between determinant, adjective and noun in terms of number
- Agreement between determinant, adjective and noun in terms of number and gender
- Reordering between noun and adjective and agreement between determinant, adjective and noun in terms of number and gender

Question

(1) Given the following rule

```
<rule>
<pattern>
<pattern-item n="determinant"/>
<pattern-item n="adjectius"/>
<pattern-item n="nom"/>
</pattern>
<action>
<out>
<lu>
<clip pos="1" side="tl" part="lem"/>
<clip pos="1" side="tl" part="a det"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
<b/>
<lu>
<clip pos="3" side="tl" part="lem"/>
<clip pos="3" side="tl" part="a nom"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
<b/>
<lu>
<clip pos="2" side="tl" part="lem"/>
<clip pos="2" side="tl" part="a adj"/>
<clip pos="3" side="tl" part="gen"/>
<clip pos="3" side="tl" part="nbr"/>
</lu>
</out>
</action>
</rule>
```

(2) Choose one interpretation

- Agreement between determinant, adjective and noun in terms of number
- Agreement between determinant, adjective and noun in terms of number and gender
- **Reordering between noun and adjective and agreement between determinant, adjective and noun in terms of number and gender**

What next?

Statistical Machine Translation