ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages

Yves LEPAGE & Etienne DENOUVAL
ATR SLC, Keihanna, Japan
Plan

- The motivation behind ALEPH: handling divergences across languages
- Q: Replicating productivity and systematicity in language?
- A: Proportional analogies!
- The translation method
- IWSLT tracks and settings
- Results and discussion
The motivation behind the ALEPH system:
- Assess the use of a unique specific operation: proportional analogy, in order to address the issue of divergences across languages in machine translation.

Features: no pattern, no variable, no training, no explicit transfer component, examples only \(\rightarrow\) a “pure” EBMT system.
Divergences

**Divergences across languages**: a study by [Dorr & al. 2002] on 19,000 sentences in English and Spanish shows that 1 sentence in 3 exhibits divergences.

- **Categorial divergence**: *tener celos (N) ↔ to be jealous (A)*
- **Conflation**: *ir flotando ↔ to float*
- **Structural divergence**: *entrar en N ↔ to enter N*
- **Head swapping**: *entrar corriendo ↔ to run in*
- **Thematic divergence**: *me gustan uvas ↔ I like grapes*
It is well known that a word does not necessarily correspond to a word.

atravesó el río flotando ↔ it floated across the river
Divergences

- It is well known that a word does not necessarily correspond to a word.

  atravesó el río flotando $\leftrightarrow$ it floated across the river

- To be able to handle divergences across languages, we need to think at a lower level than that of words, that of characters.
Productivity in language

- **Proportional analogies** are a way to achieve productivity in language (Paul, Bloomfield, Mounin, ...)

\[
\begin{align*}
\text{they swam} & \quad \text{in} \quad \text{the sea} \\
\text{they swam} & \quad \text{across} \quad \text{the river} \\
\text{it floated} & \quad \text{in} \quad \text{the sea} \\
\end{align*}
\]

\[
\begin{align*}
\text{nadaron} & \quad \text{en el} \quad \text{mar} \\
\text{atravesaron} & \quad \text{el rio} \quad \text{nadando} \\
\text{flotó en el} & \quad \text{mar} \\
\end{align*}
\]
Productivity in language

- **Proportional analogies** are a way to achieve productivity in language (Paul, Bloomfield, Mounin, ...)

- They swam
  - in the sea
- They swam
  - across the river
- It floated
  - in the sea
- It floated
  - across the river

- Nadaron
  - en el mar
- Atravesaron
  - el rio
- Nadando
- Flotó en el
  - mar
- ?
Productivity in language

- **Proportional analogies** are a way to achieve productivity in language (Paul, Bloomfield, Mounin, ...)

\[
\begin{align*}
\text{they swam} & \quad : \quad \text{it floated} \\
\text{in} & \quad : \quad \text{in} \\
\text{the sea} & \quad : \quad \text{the sea}
\end{align*}
\]

\[
\begin{align*}
\text{nadaron} & \quad : \quad \text{flotó en el} \\
\text{en el} & \quad : \quad \text{en el} \\
\text{mar} & \quad : \quad \text{mar}
\end{align*}
\]

\[
\begin{align*}
\text{they swam} & \quad : \quad \text{it floated} \\
\text{across} & \quad : \quad \text{across} \\
\text{the river} & \quad : \quad \text{the river}
\end{align*}
\]

\[
\begin{align*}
\text{atravesaron} & \quad : \quad \text{atravesó el rio} \\
\text{el rio} & \quad : \quad \text{flotando}
\end{align*}
\]

\[
\begin{align*}
\text{nadando} & \quad : \quad \text{nada en el}
\end{align*}
\]
Systematicity of language

- **Proportional analogies** are a way to make explicit the paradigmatic and syntagmatic axes of a language (Paul, Saussure, Lavie).

*It floated across the river.*
Systematicity of language

- Proportional analogies are a way to make explicit the paradigmatic and syntagmatic axes of a language (Paul, Saussure, Lavie).

It swam.

It floated.

It swam across the river.

They swam in the sea.

They swam across the river.

It walks across the street.

It walked across the street.

It floats across the river.

It floated across the river.

present / past
Systematicity of language

- **Proportional analogies** are a way to make explicit the paradigmatic and syntagmatic axes of a language (Paul, Saussure, Lavie).

*It walks across the street.* :: *It walked across the street* :: *It floats across the street.*

*It swam.* :: *It floated.* :: *It swam across the river.* :: *It floated across the river.* :: *It swam in the sea.* :: *It floated in the sea.*

*They swam across The river.* :: *They swam in The sea.* + PP or not
Systematicity of language

- **Proportional analogies** are a way to make explicit the paradigmatic and syntagmatic axes of a language (Paul, Saussure, Lavie).

\[
\begin{align*}
\text{It walks across the street.} & : \text{It walked across the street.} & : \text{It floats across the river.} \\
\text{It swam across the river.} & : \text{It floated across the river.} & : \text{It floated across the river.} \\
\text{They swam in the sea.} & : \text{They swam across the river.} & : \text{They swam across the river.}
\end{align*}
\]
Proportional analogies

- The (one and only one) technical meaning of the technical term *proportional analogies*:

- There is a *proportional analogy* between four objects A, B, C and D if and only if A is to B as C is to D.

- A proportional analogy is noted

  \[ A : B :: C : D \]
Proportional analogies

It suffices to say that we have an algorithm to verify and solve proportional analogies between strings of characters (whatever the characters) [Lepage 2003]

May I have some tea, please? : May I have a cup of coffee? :: I’d like some strong tea, please.

あの、紅茶をください。 : コーヒーをください。 :: あの、濃い紅茶がのみたい : ?
Proportional analogies

It suffices to say that we have an algorithm to verify and solve proportional analogies between strings of characters (whatever the characters) [Lepage 2003]

May I have some tea, please? : May I have a cup of coffee? :: I’d like some strong tea, please. : I’d like a cup of strong coffee.

あの、紅茶をください。 : コーヒーをください。 :: あの、濃い紅茶がのみたい : ?
It suffices to say that we have an algorithm to verify and solve proportional analogies between strings of characters (whatever the characters) [Lepage 2003]

May I have some tea, please? : May I have a cup of coffee? :: I’d like some strong tea, please. : I’d like a cup of strong coffee.

あの、紅茶をください。 : コーヒーやをください。 :: あの、濃い紅茶がのみたい : 濃いコーヒーがのみたい。
The translation method

it floated across the river

D
The translation method

A : B :: C : it floated across the river

D
The translation method

good morning : a cup of tea, please :: C : it floated across the river

buenos dias : un té, por favor :: D

The translation method

\[ \text{they swam} \quad \begin{align*} \text{in} \quad & \text{they swam} \\ \text{the sea} \end{align*} \quad :: \quad \text{across} \quad \begin{align*} \text{the river} \end{align*} \quad :: \quad \text{it floated} \quad \begin{align*} \text{across} \\ \text{the river} \end{align*} \quad \]

\[ \begin{align*} \text{nadaron} \\ \text{en el} \\ \text{mar} \end{align*} \quad \begin{align*} \text{atravesaron} \\ \text{el rio} \\ \text{nadando} \end{align*} \]
The translation method

they swam in the sea :: they swam across the river :: it floated in the sea :: it floated across the river

recursive call

nadaron en el mar :: atravesaron el rio nadando :: C :: D
The translation method

they swam in the sea :: they swam across the river :: it floated in the sea :: it floated across the river

nadaron en el mar :: atravesaron el rio nadando :: flotó en el mar :: D
The translation method

they swam in the sea : they swam across the river :: it floated in the sea : it floated across the river

nadaron en el mar : atravesaron el rio nadando :: flotó en el mar : atravesó el rio flotando
The translation method

they swam in the sea :: they swam across the river :: it floated in the sea :: it floated across the river

nadaron en el mar :: atravesaron el rio nadando :: flotó en el mar :: atravesó el rio flotando
The translation method

it floated across the river

atravesó el río flotando
Evaluation phase

- **No system tuning** or development phase
  - *Same program* used for all tracks! Only examples differ.

- Just load the training and development sets into memory.

- **No tools**: no tagger, parser, chunker, dictionaries!

- We even **REMOVED** word segmentation, in order to eliminate any influence of preprocessing tools. (or pretend there are none available on the market) (we like challenges)
Evaluation on the IWSLT2004 CE task:

<table>
<thead>
<tr>
<th>System</th>
<th>mWER</th>
<th>mPER</th>
<th>BLEU</th>
<th>NIST</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISL-S</td>
<td>0.379</td>
<td>0.319</td>
<td>0.524</td>
<td>9.56</td>
<td>0.748</td>
</tr>
<tr>
<td>ALEPH closed</td>
<td>0.434</td>
<td>0.400</td>
<td>0.522</td>
<td>8.42</td>
<td>0.687</td>
</tr>
<tr>
<td>ALEPH open</td>
<td>0.437</td>
<td>0.404</td>
<td>0.512</td>
<td>8.24</td>
<td>0.682</td>
</tr>
<tr>
<td>IRST</td>
<td>0.457</td>
<td>0.393</td>
<td>0.440</td>
<td>7.24</td>
<td>0.671</td>
</tr>
<tr>
<td>IBM</td>
<td>0.525</td>
<td>0.442</td>
<td>0.350</td>
<td>7.36</td>
<td>0.684</td>
</tr>
<tr>
<td>ISL-E</td>
<td>0.531</td>
<td>0.427</td>
<td>0.275</td>
<td>7.50</td>
<td>0.666</td>
</tr>
<tr>
<td>ISI</td>
<td>0.573</td>
<td>0.499</td>
<td>0.243</td>
<td>5.42</td>
<td>0.602</td>
</tr>
<tr>
<td>NLPR</td>
<td>0.578</td>
<td>0.531</td>
<td>0.311</td>
<td>5.92</td>
<td>0.563</td>
</tr>
<tr>
<td>HIT</td>
<td>0.594</td>
<td>0.487</td>
<td>0.243</td>
<td>6.13</td>
<td>0.611</td>
</tr>
<tr>
<td>CLIPS (systran)</td>
<td>0.658</td>
<td>0.542</td>
<td>0.162</td>
<td>6.00</td>
<td>0.584</td>
</tr>
<tr>
<td>ICT</td>
<td>0.846</td>
<td>0.765</td>
<td>0.079</td>
<td>3.64</td>
<td>0.386</td>
</tr>
</tbody>
</table>
Evaluation on the IWSLT2004 JE task:

<table>
<thead>
<tr>
<th>System</th>
<th>mWER</th>
<th>mPER</th>
<th>BLEU</th>
<th>NIST</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATR-H</td>
<td>0.263</td>
<td>0.233</td>
<td>0.630</td>
<td>10.72</td>
<td>0.796</td>
</tr>
<tr>
<td>RWTH</td>
<td>0.305</td>
<td>0.249</td>
<td>0.619</td>
<td>11.25</td>
<td>0.824</td>
</tr>
<tr>
<td>ALEPH closed</td>
<td>0.324</td>
<td>0.300</td>
<td>0.634</td>
<td>9.19</td>
<td>0.731</td>
</tr>
<tr>
<td>ALEPH open</td>
<td>0.437</td>
<td>0.403</td>
<td>0.534</td>
<td>8.97</td>
<td>0.697</td>
</tr>
<tr>
<td>U. Tokyo</td>
<td>0.485</td>
<td>0.420</td>
<td>0.397</td>
<td>7.88</td>
<td>0.672</td>
</tr>
<tr>
<td>CLIPS(systran)</td>
<td>0.730</td>
<td>0.597</td>
<td>0.132</td>
<td>5.64</td>
<td>0.568</td>
</tr>
</tbody>
</table>
Evaluation on the IWSLT2005 task:


- For the JE, CE, EC, KE tracks:
  - C-STAR Basic Traveler’s Expressions Corpus (BTEC)
    - 162,318 translation pairs

- For the AE track, only the supplied 20k set was used.

- “Open” configuration: sentences in the “training set” are left out if they appear in the task:

  **NO exact matches allowed!** The system has to re-translate.
Evaluation on the IWSLT2005 task:

<table>
<thead>
<tr>
<th>Track</th>
<th>mWER</th>
<th>mPER</th>
<th>BLEU</th>
<th>NIST</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KE</td>
<td>0.530</td>
<td>0.486</td>
<td>0.412</td>
<td>7.12</td>
<td>0.446</td>
</tr>
<tr>
<td>CE</td>
<td>0.454</td>
<td>0.418</td>
<td>0.477</td>
<td>7.85</td>
<td>0.553</td>
</tr>
<tr>
<td>JE</td>
<td>0.361</td>
<td>0.323</td>
<td>0.593</td>
<td>9.82</td>
<td>0.607</td>
</tr>
<tr>
<td>AE</td>
<td>0.527</td>
<td>0.497</td>
<td>0.382</td>
<td>6.22</td>
<td>0.481</td>
</tr>
<tr>
<td>EC</td>
<td>0.798</td>
<td>0.746</td>
<td>0.097</td>
<td>3.03</td>
<td>0.363</td>
</tr>
</tbody>
</table>
Summary

- **Method:** Use of a unique specific operation, namely *proportional analogy*, to address the issue of divergences across languages in machine translation.

- **Features:** no pattern, no variable, no training, no explicit transfer component, examples only [character strings] → a pure EBMT system.

Thank you for your attention!
Results

わかりました。

322 Did you hear me?
1 Okay.
Results

- 今、どのあたりですか。
  160  Where are we now?
  8    Here are we now?

- 私には大きすぎます。
  8257  It’s too large for me.
  8253  It’s too big.
  8251  It’s too large.
  8242  This is too large for me.
Results

航空便でいくらかかりますか。

73  How much does it cost by airmail?
56  About how much does it cost by airmail?
52  How much will it be by airmail?
20  About how much will it be by airmail?
Results

このツアーの料金はいくらですか。

271 How much does this tour cost?
160 How much do you charge for this tour?
141 What’s the price of this tour?
94 What does this tour cost?
43 What’s the price of the tour?
6 What is the price of the tour?
6 How much is the green fee?
Results

胃が痛いんです。

1744 I have a stomach ache.
552 My stomach hurts.
124 I’ve got a stomach ache.
56 Do you have a stomach ache.
51 Do you have a stomach ache?
50 I have a stomach ache?
2 My stomach hurts me.
1 I have an abdominal pain in my stomach.
1 I have a pain in my stomach.
1 I have a sore throat.