Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on `apertium.org`

Jim O'Regan[1] and Mikel L. Forcada[2]

[1] Eolaistriu Technologies, Thurles (Ireland)

[2] Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant (Spain)

SEPLN 2013, Madrid, September 18–20, 2013

Contents

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Outline

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Machine translation for Basque/1

There are two main uses for machine translation (MT)

- Dissemination: MT output is *post-edited* to produce a translation that will be published.
- Assimilation or gisting: MT output is used *as is* to understand text written in another language

Machine translation for Basque/2

Unlike other languages, the Basque language has no living *cousins*: it is hard to understand for almost everyone else.

Assimilation MT systems for Basque are useful for those wanting to follow Basque affairs.

Machine translation for Basque/3

Why free/open-source MT from Basque?

- Basque is supported, for instance, by Google. However:
  - Google is statistical MT and sometimes favours *fluency* over *adequacy* (='fidelity') [example: missing *don't*]
  - Google is online: users may not want confidential or sensitive data to travel there and back
  - The resources used by Google are not available for other applications
- Having free/open-source rule-based MT from Basque:
  - ensures that adequacy is preserved (perhaps at the expense of fluency)
  - makes linguistic resources (dictionaries, rules) available to a wider community (to create new NLP applications)
  - allows for offline usage on sensitive material

Outline

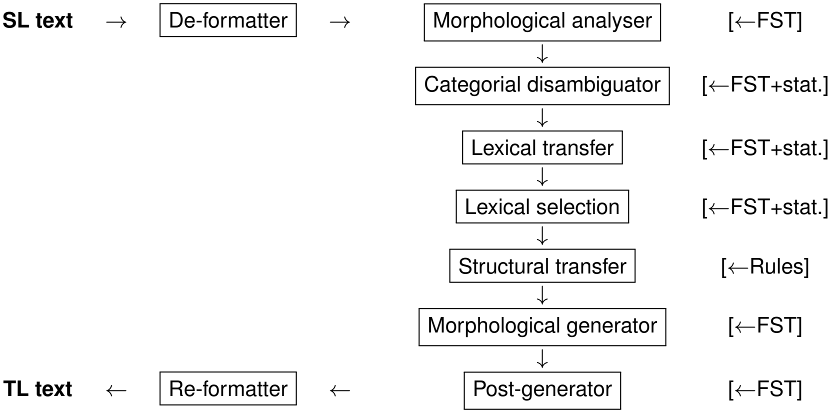
1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

The Apertium MT platform/1

Apertium is a free/open-source machine translation platform (`http://www.apertium.org`) providing:

1. A free/open-source modular shallow-transfer machine translation **engine** with:
   - text format management
   - finite-state lexical processing and lexical selection
   - statistical (HMM) and rule-based (CG) lexical disambiguation
   - shallow transfer based on finite-state pattern matching
2. Free/open-source **linguistic data** in well-specified XML formats for a variety of language pairs (35 *stable* pairs)
3. Free/open-source **tools**: compilers to turn linguistic data into the fast and compact form used by the engine, software to learn disambiguation or transfer rules, etc.

The Apertium MT platform/2



SL text → De-formatter → Morphological analyser [←FST]

↓

Categorial disambiguator [←FST+stat.]

↓

Lexical transfer [←FST+stat.]

↓

Lexical selection [←FST+stat.]

↓

Structural transfer [←Rules]

↓

Morphological generator [←FST]

↓

TL text ← Re-formatter ← Post-generator [←FST]

The Apertium MT platform/3

Communication between modules: text (Unix “*pipelines*”).
Advantages:

- Simplifies diagnosis and debugging
- Allows the modification of data between two modules using, e.g., filters
- Makes it easy to insert alternative modules (interesting for research and development purposes)
    - An example: some language pairs have an alternative finite-state processor for morphological analysis and generation (based on HFST).

Outline

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Apertium Basque to English /1

We were able to reuse existing data:

- Basque morphological dictionary from **`apertium-eu-es`** (Ginestí-Rosell et al. 2011), most coming from Matxin (Mayor et al. 2011).
- English morphological dictionary from **`apertium-en-es`**
- Bilingual dictionary obtained by crossing the bilingual dictionaries in **`apertium-eu-es`** and **`apertium-en-es`** using **`apertium-dixtools`** and manually extending, aided with existing English–Basque data in Matxin.
- Basque part-of-speech tagger from **`apertium-eu-es`**
- Structural transfer rules: adapted from **`apertium-eu-es`** and extended (noun–noun compounds, verbs, dates)

Apertium Basque to English /2

- The data were then manually corrected and completed
- Brief description of the data (rev. 36906):

ITEM	COUNT
Number of Basque→English dictionary entries	9, 594
Total structural transfer rules	272

Outline

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Evaluation of gisting: a strategy/1

- Evaluating MT for gisting or assimilation is not easy.
- Standard approaches use a costly "reading comprehension" approach with carefully-crafted TL questions (Jones et al. 2007)
- Alternative methods based on blind post-editing followed by human assessment of adequacy are also expensive (WMT 2009, 2010; Ginestí-Rosell et al. 2009).
- We want a less expensive way to evaluate how much MT improves understanding of *foreign* text.
- We have devised a novel *cloze test* (*closure test*) strategy, starting with a parallel corpus
  - Cloze tests have so far been performed on raw MT output, not on reference sentences (Somers and Wild 2000).

Evaluation of gisting: a strategy/2

The procedure:

- Create *holes* or *gaps* in the reference target-language (TL) sentences by randomly blanking out a certain fraction (e.g. 20%) of content words (i.e., not stop-words)
    - Blanked-out words marked by a placeholder, e.g. #####
- Ask non-TL-speaking subjects to complete randomly chosen TL sentences in 4 different *hinting* situations:
    - Without any hint whatsoever
    - Showing the SL sentence (expected to help little)
    - Showing the TL sentence produced by MT
    - Showing both

Evaluation of gisting: a strategy/3

An example:

Basque (source language) hint:	`Bruselako Adierazpenaren sinatzaileek argi eta garbi zuzendu dute Adierazpen horrek ordezkatzen duen nazioarteko komunitatearen eskaera.`
Machine translation hint:	`the signatories of the Statement of Brussels clear and clean they have addressed this Statement he of the international community that substitutes the request.`
Problem sentence:	[sm]@@143: The **#####** of the **#####** Declaration have addressed in **#####** **#####** the demands of the international **#####** which the **#####** Declaration represents.
Reference sentence:	The **endorsers** of the **Brussels** Declaration have addressed in **unequivocal terms** the demands of the international **community** which the **Brussels** Declaration represents.

Evaluation of gisting: a strategy/4

“Synonyms” may (optionally) be allowed:

mesures	measures
mandate	Mandate
likewise	also
legalization	legalisation
lawful	legitimate
laid	set
kept	maintained
international	International
HNT	ICG
…	…

Outline

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Results/1

Experimental settings:

- 23 (out of 27) informants with good English command and no command of Basque
- 20% content words blanked out (well beyond the monolingual guessing threshold)
- each informant received 32 problems
- roughly 8 of each kind (no hint, SL hint, MT hint, both)
- **`apertium-eu-en`** rev. 39606
- 86-entry synonym list conservatively built by inspecting valid alternatives (after informants finished their work).

Results /2

HINT LEVEL	# OF 1-WORD HOLES	SUCCESS RATE (EXACT)	SUCCESS RATE (SYNONYMS)
No hint	575	26% (sd 13%)	30% (sd 14%)
SL hint	543	29% (sd 12%)	34% (sd 14%)
MT hint	597	48% (sd 13%)	54% (sd 13%)
Both hints	589	43% (sd 13%)	51% (sd 14%)

- Success is high with no hints (repetitive, predictable text)
- SL hint not too useful (proper nouns and cognates?)
- Success rate improves clearly with MT hint
- Having both hints seems to hurt
- Synonyms do not change general trend

Outline

1. Machine translation for Basque
2. The Apertium MT platform
3. Apertium Basque to English
4. Evaluation of gisting: a novel strategy
5. Results
6. Conclusions and future work

Main contributions

Two main contributions in this (very preliminary) work:

- A working prototype of a free/open-source rule-based MT system from Basque to English: **`apertium-eu-en`**
- A new, cheap method to evaluate the usefulness of an MT system for *gisting* based on *cloze* (*closure*) tests.

Conclusions

Main findings:

- It is possible to build, in a few months, a Basque-to-English MT system capable of improving the level of understanding, on the part of non-Basque speakers, of the contents of Basque test.
- A simple, inexpensive method may be used to assess this improvement.

Future work

- Improving **`apertium-eu-en`** further:
  - For assimilation purposes
  - For interactive *predictive* translation “à la” Transtype
- Performing a more extensive evaluation:
  - Using a non-repetitive corpus to minimize “monolingual guessing”
  - Studying the effect of the percentage of gaps
  - Using other MT systems (such as Google Translate) as MT hints to perform a comparative evaluation
  - Studying the correlation with more expensive evaluations of gisting.

Acknowledgements

The authors thank

- the European Association for Machine Translation (EAMT) for financial support
- all 23 informants for participating in the evaluation,
- Francis M. Tyers for useful discussions on the idea of using Cloze tests for machine translation evaluation, and
- Mireia Ginestí for her help on questions regarding Basque data.

License