Tapadóir: Statistical Machine Translation for Irish

Meghan Dowling

ADAPT Centre, School of Computing, Dublin City University, Ireland

29th April 2015
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
Motivation for implementing a SMT system

- Current demand for translated content exceeds the capacity of the DAHG’s translation team
- Previously:
  - SDL Trados (TM)
  - Not suitable for previously unseen text
  - If low % TM match, must be translated from scratch
- Clear need for something which would speed up the translator’s work, while also being domain-specific
Motivation for implementing a SMT system

- Current demand for translated content exceeds the capacity of the DAHG’s translation team
- Previously:
  - SDL Trados (TM)
  - Not suitable for previously unseen text
  - If low % TM match, must be translated from scratch
- Clear need for something which would speed up the translator’s work, while also being domain-specific
Motivation for implementing a SMT system

- Current demand for translated content exceeds the capacity of the DAHG’s translation team
- Previously:
  - SDL Trados (TM)
  - Not suitable for previously unseen text
  - If low % TM match, must be translated from scratch
- Clear need for something which would speed up the translator’s work, while also being domain-specific
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
The Tapadóir Project

Main goals:

▶ Build a SMT engine tailored to the translation needs of the DAHG
  ▶ Domain-specific
  ▶ User-friendly and non-disruptive to the usual workflow
▶ Gather and curate appropriate data
▶ Explore algorithmic parameters to best fit the language(s) and use case
The Tapadóir Project

Main goals:

- Build a SMT engine tailored to the translation needs of the DAHG
  - Domain-specific
  - User-friendly and non-disruptive to the usual workflow
- Gather and curate appropriate data
- Explore algorithmic parameters to best fit the language(s) and use case
The Tapadóir Project

Main goals:

- Build a SMT engine tailored to the translation needs of the DAHG
  - Domain-specific
  - User-friendly and non-disruptive to the usual workflow
- Gather and curate appropriate data
- Explore algorithmic parameters to best fit the language(s) and use case
Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate

▶ reports, staff notices, communications, annual reports, etc
▶ a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement
Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate

- reports, staff notices, communications, annual reports, etc
- a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement.
Defining a Use-Case

During the Pilot phase, translators were consulted to identify the type of text Tapadóir will be used to translate:

- reports, staff notices, communications, annual reports, etc
- a formal tone and at times, a high register

This also established a translator feedback loop, which is now used for identifying areas for improvement.
Acquiring Data

- Previous publicly available corpora:
  - Parallel English–Irish corpus of legal texts (Paradocs): English-Irish corpus of legal texts
  - Corpas Comhthreomhar Gaeilge-Béarla (CCGB): Bilingual corpus crawled from the web

- Corpora gathered by Tapadóir:
  - General domain data: Data crawled using the ILSP web-crawler
  - Domain-specific data: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

---

1 [www.gaois.ie](http://www.gaois.ie)
2 [http://borel.slu.edu/corpas/](http://borel.slu.edu/corpas/)
Acquiring Data

- Previous publicly available corpora:
  - **Parallel English–Irish corpus of legal texts (Paradocs)**\(^1\): English-Irish corpus of legal texts
  - **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**\(^2\): Bilingual corpus crawled from the web

- Corpora gathered by Tapadóir:
  - **General domain data**: Data crawled using the ILSP web-crawler\(^3\)
  - **Domain-specific data**: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

\(^1\)www.gaois.ie
\(^2\)http://borel.slu.edu/corpas/
\(^3\)http://nlp.ilsp.gr/redmine/projects/ilsp-fc
Acquiring Data

- Previous publicly available corpora:
  - **Parallel English–Irish corpus of legal texts (Paradocs)**\(^1\): English-Irish corpus of legal texts
  - **Corpas Comhthreomhar Gaeilge-Béarla (CCGB)**\(^2\): Bilingual corpus crawled from the web

- Corpora gathered by Tapadóir:
  - **General domain data**: Data crawled using the ILSP web-crawler\(^3\)
  - **Domain-specific data**: Translation memories received from the Dept. of Arts, Heritage and the Gaeltacht and the Joint Research Centre of the European Commission

---

\(^1\)www.gaois.ie
\(^2\)http://borel.slu.edu/corpas/
\(^3\)http://nlp.ilsp.gr/redmine/projects/ilsp-fc
Increased bilingual corpora

Table of Irish-English corpora gathered

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size(lines)</th>
<th>Size(words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paradocs</td>
<td>89,000</td>
<td>1,526,498</td>
</tr>
<tr>
<td>CCGB</td>
<td>6,000</td>
<td>113,889</td>
</tr>
<tr>
<td>DAHG (baseline)</td>
<td>29,000</td>
<td>67,418</td>
</tr>
<tr>
<td>DAHG (Additional)</td>
<td>13,500</td>
<td>68,691</td>
</tr>
<tr>
<td>Crawled (cleaned)</td>
<td>10,000</td>
<td>183,999</td>
</tr>
<tr>
<td>Crawled (uncleaned)</td>
<td>55,000</td>
<td>1,062,942</td>
</tr>
<tr>
<td>DCEP &amp; DGT-TM</td>
<td>29,000</td>
<td>439,262</td>
</tr>
</tbody>
</table>

Table 1: Current data sets collected for Irish↔English MT. Word counts given for the English files only.
Developing a test set

- 1500 sentences of gold-standard TM data provided by DAHG
- Domain-specific
- Representative of the type of text Tapadóir will need to translate
- Used to evaluate using automatic metrics
SMT Engine

- Built using Moses 2.0 (Baseline) and Moses 3.0 (Current)
- 6-gram language model
- Trained using combinations of previously mentioned corpora
- KenLM
System Setting Modifications

▶ Tuning
  ▶ Performed on a held-out section of 3000 domain-specific sentence pairs

▶ Hierarchical-based model
  ▶ Reordering table changed from phrase-based to hierarchical
  ▶ Better able to handle larger ordering differences – treats adjacent pairs as one unit
System Setting Modifications

- **Tuning**
  - Performed on a held-out section of 3000 domain-specific sentence pairs

- **Hierarchical-based model**
  - Reordering table changed from phrase-based to hierarchical
  - Better able to handle larger ordering differences – treats adjacent pairs as one unit
Post-Processing Module

Automated Post-Editing

- corrects common MT mistakes
- uses Irish surface orthography rather than deeper morphological analysis
- aims to improve grammar and readability
- \textit{ag mé*} $\rightarrow$ \textit{agam}
  - \textit{ag an baile*} $\rightarrow$ \textit{ag an mbaile}
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
## System Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>33.91</td>
</tr>
<tr>
<td>Tapadóir (Baseline)</td>
<td>39.44</td>
</tr>
<tr>
<td>Tapadóir (Current)</td>
<td>43.19</td>
</tr>
</tbody>
</table>

**Table 2**: Comparison of results for Google Translate, our baseline system and our current system
Increased MT usage
Integration into DAHG translation workflow
Integration into DAHG translation workflow

Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said:

“This promises to be an extremely exciting and uplifting event; the ideal start to our year of commemorations in 2016.

The Peace Proms will be a wonderful occasion, bringing together talented and committed young musicians from across the island of Ireland.

I would like to thank all of the young performers who will be taking part; through their music and song they will set the perfect tone as kick off what is set to be a very special year.”
Integration into DAHG translation workflow
Typical MT errors

- Incorrect case used
- Inaccurate reordering - especially with longer sentences
- Copula constructions
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
Ongoing and Future Work

- Gather additional general-domain parallel data

- **Source-side re-ordering**
  To address divergent word order (SVO–VSO)

  **SRC:** The timeframe *can* be extended
  **RO:** *can* The timeframe be extended
  **REF:** ‘*Is féidir an tráthchlár a shíneadh*’

- **Factored Models**
  Using a POS-tagger (Úi Dhonnchadha), build a language model which contains grammatical information
Ongoing and Future Work

- Gather additional general-domain parallel data
- **Source-side re-ordering**
  To address divergent word order (SVO–VSO)
  
  **SRC:** The timeframe *can* be extended
  **RO:** *can* The timeframe be extended
  **REF:** ‘Is féidir an tráthchlár a shíneadh’

- **Factored Models**
  Using a POS-tagger (Úí Dhonnchadha), build a language model which contains grammatical information
Ongoing and Future Work

- Gather additional general-domain parallel data
- **Source-side re-ordering**
  To address divergent word order (SVO–VSO)
  
  **SRC:** The timeframe *can* be extended
  **RO:** *can* The timeframe be extended
  **REF:** ‘*Is féidir* an tráthchlár a shíneadh’

- **Factored Models**
  Using a POS-tagger (Úi Dhonnchadh), build a language model which contains grammatical information
Ongoing and Future Work

- DAHG vision for a Shared Translation Service
- Interest from DGT
- ELRC helping data gathering cause
  - DAHG helping to identify and gather public admin data from around the country
Outline

Motivation

The Tapadóir Project

Results

Ongoing and Future Work

Conclusion
Conclusion

- Clear need for further research in English-Irish MT
- Clear need for new methods in overcoming the morphology problem/ lack of data problem
Conclusion

Go Raibh Maith Agaibh!
(Thank You!)