Telemedicine as a special translation

Author: Krzysztof Wołk
Importance

• Narrow text domain
• Important and promising field of research
• Medical records translation
• Foreign medical history access
• Communication between patient and doctor abroad
• Telemedicine
• Lower costs that human translators
The idea ...

<table>
<thead>
<tr>
<th>Domain</th>
<th>WER</th>
<th>Vocabulary size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polish Senate</td>
<td>19.6</td>
<td>87k</td>
</tr>
<tr>
<td>TV news</td>
<td>15.72</td>
<td>42k</td>
</tr>
<tr>
<td>Lectures</td>
<td>27.75</td>
<td>210k</td>
</tr>
</tbody>
</table>
EU-BRIDGE Project
Requirements

- Bilingual Parallel Data
- Alignment
- Monolingual Data for Language Model
- Data in good quality
- Data that is natural
- Really Lots of data
Problems with SMT for PL – EN Pair

- Not Enough Polish Data
- Poor Quality of Accessible Polish Data
- Bad, poor or too indirect translations

Sentence 1: ABCD  Sentence 2: WXYZ

ABABCD.  ABCBCD.

ABCDABCD. ABWXYZCD.

- E.g. about 10% spelling 17% insertions 6% translations in TED
Differences in languages

- Different syntactic order (SVO)
  - I bought myself a new car.
  - Kupiłem sobie nowy samochód.
  - Nowy samochód sobie kupiłem.
  - Sobie kupiłem nowy samochód.
  - Samochód nowy sobie kupiłem.

- Bigger Vocabulary (E.g. TED, PL: 92135, EN: 41684)
- More complex grammar

  Cases:
  - PL: 7 vs EN: 3

In Polish 15 gender forms for nouns and adjectives, with additional dimensions for other word classes.
The EMEA Corpora

European Medical Agency (EMEA)
Leaflets from the drugs + Documents

Corpora size: about 80 MB

1,044,764 bi-sentences

11.67M words that were not tokenized

148,170 unique PL words
109,326 unique EN words
TOOLS

- Moses
  - Pipeline
  - Decoder
- Wroclaw NLP
- JMX Tagger
- EMS
Automatic Evaluation Methods for MT

- **BLEU** – Bilingual Evaluation Understudy
- **NIST** – National Institute of Standards and Technology
- **METEOR** – Metric for Evaluation of Translation with Explicit Ordering
- **TER**
## RESULTS

### Polish-to-English translation

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>70.15</td>
<td>10.53</td>
<td>82.19</td>
<td>29.38</td>
</tr>
<tr>
<td>01</td>
<td>64.58</td>
<td>9.77</td>
<td>76.04</td>
<td>35.62</td>
</tr>
<tr>
<td>02</td>
<td>71.04</td>
<td>10.61</td>
<td>82.54</td>
<td>28.33</td>
</tr>
<tr>
<td>03</td>
<td>71.22</td>
<td>10.58</td>
<td>82.39</td>
<td>28.51</td>
</tr>
<tr>
<td>04</td>
<td>76.34</td>
<td>10.99</td>
<td>85.17</td>
<td>24.77</td>
</tr>
<tr>
<td>05</td>
<td>70.33</td>
<td>10.55</td>
<td>82.28</td>
<td>29.27</td>
</tr>
<tr>
<td>06</td>
<td>71.43</td>
<td>10.60</td>
<td>82.89</td>
<td>28.73</td>
</tr>
<tr>
<td>07</td>
<td>71.91</td>
<td>10.76</td>
<td>83.63</td>
<td>26.60</td>
</tr>
<tr>
<td>08</td>
<td>71.12</td>
<td>10.37</td>
<td>84.55</td>
<td>29.95</td>
</tr>
<tr>
<td>09</td>
<td>71.32</td>
<td>10.70</td>
<td>83.31</td>
<td>27.68</td>
</tr>
<tr>
<td>10</td>
<td>71.35</td>
<td>10.40</td>
<td>81.52</td>
<td>29.74</td>
</tr>
<tr>
<td>11</td>
<td>70.34</td>
<td>10.64</td>
<td>82.65</td>
<td>28.22</td>
</tr>
<tr>
<td>12</td>
<td>72.51</td>
<td>10.70</td>
<td>82.81</td>
<td>28.19</td>
</tr>
</tbody>
</table>
# Results

## English-to-Polish translation

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>69.18</td>
<td>10.14</td>
<td>79.21</td>
<td>30.39</td>
</tr>
<tr>
<td>01</td>
<td>61.15</td>
<td>9.19</td>
<td>71.91</td>
<td>39.45</td>
</tr>
<tr>
<td>02</td>
<td>69.41</td>
<td>10.14</td>
<td>78.98</td>
<td>30.90</td>
</tr>
<tr>
<td>03</td>
<td>68.45</td>
<td>10.06</td>
<td>78.63</td>
<td>31.62</td>
</tr>
<tr>
<td>04</td>
<td>73.32</td>
<td>10.48</td>
<td>81.72</td>
<td>27.05</td>
</tr>
<tr>
<td>05</td>
<td>69.21</td>
<td>10.15</td>
<td>79.26</td>
<td>30.88</td>
</tr>
<tr>
<td>06</td>
<td>69.27</td>
<td>10.16</td>
<td>79.30</td>
<td>31.27</td>
</tr>
<tr>
<td>07</td>
<td>68.43</td>
<td>10.07</td>
<td>78.95</td>
<td>33.05</td>
</tr>
<tr>
<td>08</td>
<td>67.61</td>
<td>9.87</td>
<td>77.82</td>
<td>29.95</td>
</tr>
<tr>
<td>09</td>
<td>68.98</td>
<td>10.11</td>
<td>78.90</td>
<td>31.13</td>
</tr>
<tr>
<td>10</td>
<td>68.67</td>
<td>10.02</td>
<td>78.55</td>
<td>31.92</td>
</tr>
<tr>
<td>11</td>
<td>69.01</td>
<td>10.14</td>
<td>79.13</td>
<td>30.84</td>
</tr>
<tr>
<td>12</td>
<td>67.47</td>
<td>9.89</td>
<td>77.65</td>
<td>33.32</td>
</tr>
</tbody>
</table>
Conclusions

• Hierarchical Models – biggest improvement
• Truecasing & Punctuation Normalization
• PL -> EN better than EN -> PL
• High BLEU score (over 70)
• Beneficial to the patients
• Aid for human interpreters
TRANSLATION EXAMPLES O-ORIGINAL T-TRANSLATED

O: I think we should reflect carefully before we do that again.
T: I believe that we should reflect carefully before we do that again.

O: International cooperation is an essential tool for an effective fight against this scourge, at both national and multilateral levels.
T: International cooperation is an essential tool for an effective fight against this scourge, at both national and international levels.

O: We have completely overhauled our policy style and the way in which we prepare and review legislation.
T: We have completely overhauled our policy style and the way in which we prepare and review legislation.

O: Abilify is a medicine containing the active substance—aripiprazole.
T: Abilify is a medicine containing the active substance aripiprazole.
O: As far as I am aware, a rather successful long term social dialogue has been conducted on these issues and, in my opinion, it is thus necessary to reconcile seemingly contradictory procedures.

T: As far as I know, in this regard have been carried out rather successful long-term social dialogue and, in my view, there is a need to reconcile seemingly contradictory procedures.

O: They are concerned with the creation of a milk fund that will support reforming activities in this sector, such as the promotion of milk consumption, including consumption in schools, or support for milk production in mountainous areas.

T: They are concerned with the creation of a milk fund that will support reforming activities in this sector, such as the promotion of milk consumption, including consumption in schools, or support for milk production in mountain areas.
**O:** Split Influenza virus, inactivated, containing antigens equivalent to the following strains:

**T:** Split influenza virus, inactivated, containing antigens equivalent to the following strains:

**O:** The immunological response may be diminished if the patient is undergoing immunosuppressant treatment.

**T:** The immunological response may be *wakened* if the patient is undergoing immunosuppressant treatment.

**O:** Optisulin was originally studied in 10 studies, in both type 1 and type 2 diabetes.

**T:** Optisulin was originally studied in 10 studies *in diabetes type 1* and in type 2.
Future Work

• Combining all experiments
• Extending training data and language models
• Stems instead of surface forms
• Merging with TTS and ASR systems
• Filtering and cleaning parallel data
• Changing word order to meet SVO order
THANK YOU FOR ATTENTION!