MACHINE TRANSLATION FOR SUBTITLES AND BEYOND

Martin Volk
Universität Zürich
volk@cl.uzh.ch
Subtitles – Audiovisual Translation

- Top of the list. Repton. That the one?
- That’s it. Left a year ago.

- Oben auf der Liste. Repton-Schule. Richtig?
- Ja. Ich ging vor einem Jahr ab.

Subtitles
(by Díaz Cintas and Remael, 2007)

- **Intralingual** (same language as audio)
  - for the deaf and hard-of-hearing
  - for language learning purposes
  - for Karaoke effect
  - for dialects
  - for notices and announcements

- **Interlingual** (different language than audio)
  - for hearers
  - for the deaf and hard-of-hearing
Examples of Interlingual Subtitles

<table>
<thead>
<tr>
<th>EN</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>He's not picking up.</td>
<td>Er nimmt nicht ab.</td>
</tr>
<tr>
<td>So in order to save 100 bucks, you run the risk of permanent back injury.</td>
<td>Für $100 riskierst du also bleibende Rückenschäden.</td>
</tr>
<tr>
<td>- You’re getting married?</td>
<td>- Ihr werdet heiraten?</td>
</tr>
<tr>
<td>- When did you decide?</td>
<td>- Wann habt ihr euch entschieden?</td>
</tr>
<tr>
<td>Look, I’ve been under a lot of stress lately…</td>
<td>Ja, ich hatte in letzter Zeit viel Stress</td>
</tr>
<tr>
<td>… and I’ve been releasing some tension, that’s all.</td>
<td>und ich musste etwas von dem Druck loswerden, das ist alles.</td>
</tr>
</tbody>
</table>

Media Adaptation / Translation

- Subtitling
  - Netherlands, Portugal, Scandinavia, …
- Dubbing
  - France, Germany, Italy, Spain, …
- Voice Over
  - Eastern Europe, …
- Related to: Intertitling, Surtitling, Video Game Localization, Audio Description, …
Subtitling

- Different types of production:
  - Live or Real-time produced subtitles
    - Human made \(\rightarrow\) respeaking
    - Automatic
  - Pre-produced subtitles

- Coverage
  - Full text / dialogue subtitles
  - Reduced subtitles

Automatic Translation of TV Subtitles

- a project by
  - our Computational Linguistics group together with
  - a large subtitling company

- 2006 – 2010

- language pairs:
  - Swedish \(\rightarrow\) Danish
  - Swedish \(\rightarrow\) Norwegian
  - English \(\rightarrow\) Swedish
### English
- What's that all about?
- I can't wait for this day to be over.
- Thank God it's only once a year.
- Maybe I'll meet him tomorrow.
- Chris wants to start up a band.

### German
- Was soll denn das?
- Ich bin so froh, wenn dieser Tag vorbei ist.
- Gott sei Dank ist er nur einmal im Jahr.
- Vielleicht werde ich ihn morgen treffen.
- Chris möchte eine Band gründen.

### Statistical Machine Translation
- What's that all about?
- I can't wait for this day to be over.
- Thank God it's only once a year.
- Maybe I'll meet him tomorrow.
- Chris wants to start up a band.
What's that all about? I can't wait for this day to be over. Thank God it's only once a year. Maybe I'll meet him tomorrow. Chris wants to start up a band.

Maybe I'll start up a band.

Three Favorable Conditions for Automatic Translation

1. Subtitles are short.
   - Limited to two rows times 37 characters.
   - Average: ~ 10 words/subtitle
2. Closely related languages.

Unfavorable Conditions for Automatic Translation

- TV subtitles come from a large diversity of textual domains.
  - Comedy, Crime, Documentary, Talk show, ...
- TV subtitles are a sort of (abbreviated) transcription of spoken language.
  - Emphatic spelling, Spacing, Bold-face
    - lo-o-ove, Jerry S, e, i, n, f, e, l, d
  - Many contractions
    - Nicole’s right. We’d appreciate your help.
Evaluation Scenario

Evaluation Results against a Prior Human Translation
(for Swedish → Danish)
Evaluation Results averaged over 6 Post-editors
(for Swedish → Danish)

Evaluation Results in BLEU scores
Interview with Post-editors

- interviews with two post-editors
- generally positive impression
- post-editing is different from translating
- post-editing is hard cognitive work
  - danger of being lead in a particular direction
- estimated time-saving: 20-30%
- big differences between post-editors (generous vs. picky).

Martin Volk, Uni Zurich  1. April 2011
Time Saving Study

- by Sheila C. M. de Sousa, Wilker Aziz and Lucia Specia (University of Wolverhampton, 2011)
- English → Portuguese
- Google Translate vs. Own MT vs. Translation from Scratch
- Large experiment with 11 “volunteers”
  - native speakers of Brazilian Portuguese
  - “some experience with translation tasks”
  → Postediting is 40% faster than Translation from Scratch!

Subtitle Translation vs. Sentence Translation

Master Project by Susanne Rozkosny at ZHAW
- with supervision by Gary Massey

- Subtitle-based MT (EN → DE)
  - Look, I've been under a lot of stress lately...
  - … and I've been releasing some tension, that’s all.

- Sentence-based MT (EN → DE)
  - Look, I've been under a lot of stress lately and I've been releasing some tension, that’s all.

  → Better quality for subtitle-based MT!
EU-Project: www.sumat-project.eu

- coordinated by VicomTech, San Sebastian, Spain
- with technical partners in Greece, Ireland, Slovenia, Switzerland
- with subtitling companies
  - Deluxe Digital Studios
  - Invision Ondertiteling
  - Titelbild Subtitling and Translation
  - Voice & Script Int. (VSI)
- 2011-2014
- EN - DE, ES, FR, NL, PT, SV plus Slovenian and Serbian

English – Swedish MT in SUMAT

Parallel Subtitles as Training data
- SUMAT data (contributed by the subtitle companies)
  - 548,000 EN – SV subtitle pairs
  - 5.6 million EN tokens / 4.6 million SV tokens
- Open Subtitles (freely available on the web)
  - 7.2 million EN – SV subtitle pairs
  - 55.3 million EN tokens / 48.1 million SV tokens

Monolingual data for the Language Model
- English: 8.1 million subtitles / 76.3 million tokens
- Swedish: 3.7 million subtitles / 37.8 million tokens
Quality Estimation

- = Confidence Estimation
- = The SMT system delivers only translations when confident, otherwise nothing.

- Quality Estimation is trained as a classifier based on features such as
  - subtitle length
  - number of unknown words
  - number of translation options
  - ...

20-Mar-14
Conclusions

- Subtitling is an interesting field for translators.
- Subtitles are well-suited for Statistical Machine Translation.
- Statistical Machine Translation enables you to re-use your translation archive and to profit from previous work.

Domain-specific SMT

Goal: SMT systems for Alpine texts (DE-FR)

Based on the Text+Berg corpus

- SAC texts from 1864 – 2011
- Parallel part 1957 – 2011: ~6 million tokens
Domain-specific SMT

- **Source**: 20. Juni: unser dritter Angriff auf das Gross Grünhorn (4044 m).
  - Human: Le 20 juin eut lieu notre troisième tentative au Gross Grünhorn (4044 m).
- **Google**: 20. Juin : notre troisième attaque majeure sur la Corne de Green (4044 m).
- **We**: 20 juin : notre troisième tentative vers le Gross Grünhorn (4044 m).

---

Domain-specific SMT: Research Questions

1. How can we combine **large out-of-domain corpora** with small domain-specific corpora?
2. How can we combine **large monolingual corpora** on the source or target side with small domain-specific corpora?
3. How can we use domain-specific **terminology** to improve an SMT system?
Domain-specific SMT: Results

<table>
<thead>
<tr>
<th></th>
<th>German → French</th>
<th>BLEU</th>
<th>French → German</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Google Translate:</td>
<td>12.95</td>
<td>Google Translate:</td>
<td>10.53</td>
</tr>
<tr>
<td>2</td>
<td>Personal Translator 14:</td>
<td>13.29</td>
<td>Personal Translator 14:</td>
<td>08.93</td>
</tr>
<tr>
<td>3</td>
<td>Moses (Europarl data):</td>
<td>11.46</td>
<td>Moses (Europarl data):</td>
<td>09.27</td>
</tr>
<tr>
<td>5</td>
<td>+ BleuAlign:</td>
<td>18.18</td>
<td>+ BleuAlign:</td>
<td>15.47</td>
</tr>
<tr>
<td>6</td>
<td>+ Reordering:</td>
<td>18.38</td>
<td>+ Reordering:</td>
<td>15.69</td>
</tr>
<tr>
<td>7</td>
<td>+ Lattice:</td>
<td>18.80</td>
<td>+ Lattice:</td>
<td>15.81</td>
</tr>
<tr>
<td>8</td>
<td>+ Interpolated LM</td>
<td>18.98</td>
<td>+ Interpolated LM</td>
<td>16.69</td>
</tr>
<tr>
<td>9</td>
<td>+ Combined Translation Model</td>
<td>19.13</td>
<td>+ Combined Translation Model</td>
<td>16.69</td>
</tr>
<tr>
<td>10</td>
<td>+ Dynamic Translation Model</td>
<td>20.24</td>
<td>+ Dynamic Translation Model</td>
<td>16.69</td>
</tr>
</tbody>
</table>

Hybrid Machine Translation for Lesser-resourced languages

- We are investigating
  Spanish → Quechua Machine Translation
- Funded by SNF 2011-2015
Entonces, desde ese día, en mi corazón se prendió, como aífer, la idea de salir de la casa de mi madrina para ir buscar trabajo.

Ojalá esta señora de buen corazón el Señor la haya hecho sentir así, porque ella es la que me salvo de lo que ya estaba caminando de otra vida.

En lo que estaba caminando allá tras las ovejas, con el corazón puesto ya en otro pueblo, un día pasaron unos ancianos con dirección a Acopía.

Así, en lo que estaba caminando entre las ovejas, llorando yendo mi sueño amarga como la sal, una mujer de buen corazón me levó: Ya no lloro, hace mucho rato se han ido, me dijo.

Chayqa chay p'unchamanta pintía mami, tupa simiyuy t'ipanayach'uyuyi, sunquyacu aqua llina t'apayach'uyuyi t'ipanayach'uyuy irayach'uyuy.

Ojalá chay sunquyuq, sunquyoq soñonata Taytanchis ladonpi lluyach'urey, payni sahawaran t'a huq kax, viwamant punisqaych'uy.

Chhaywa uwiqunaq qhipan punisqaych'yu, t'a huq lado f'acapi, sunquyoq, huk p'unchay amirokuna p'acapi yuych'uy anochi laman.

Chayqa uye ukipu kachi vidsaymanta takaika wakaych'uy purushaych, huk allin sunquyucu warmi putawaw.

20-Mar-14
**MT for the Automobile domain**
*(KTI project 2012-2013)*

- Partner: Translation Company Semiotic Transfer (Baden, AG)
- Languages: DE → FR and DE → IT
- Texts: Marketing Material for Automobile Industry

---

**MT for Sign Language**

- Translation: German → Swiss German Sign Language
- via Sign Language Notation + Avatar
- Domain: Railway Announcements

Example: “Dieser Zug endet in Brig”

→ → →

→ → →

20.03.14
Autour du Bietschhorn.

Juhel 1957.

Par Ed. Jeanneau-Pernet (Société de la Chaise du Rhône).

Notre projet, cette année, était de recenser quelques jours de vacances à l'exploration des chaînes de montagnes se situant au Bietschhorn; ces chaînes, délimitées tout à l'ouest du massif du mont Blanc, occupant encore à l'heure actuelle un rôle presque virginal. Elles attireront plus tard un courant de navigation; leur base, à des profondeurs aussi considérables, doit être choisie avec soin. Les chaînes de montagnes occupent une place importante dans la géographie de la vallée. Elles forment une barrière naturelle qui sépare la Vallée du Rhône des Alpes du Nord.

La Vallée du Rhône, longue et étroite, est entourée de montagnes massives. Les chaînes de montagnes se prolongent du sud au nord, formant une barrière naturelle qui sépare la Vallée du Rhône des Alpes du Nord. Les chaînes de montagnes occupent une place importante dans la géographie de la vallée. Elles forment une barrière naturelle qui sépare la Vallée du Rhône des Alpes du Nord.

La situation géographique de la Vallée du Rhône est particulièrement intéressante. Elle se trouve à la confluence de plusieurs cours d'eau, ce qui contribue à l'irrigation et à l'agriculture. Les chaînes de montagnes qui entourent la vallée offrent des conditions idéales pour le développement de l'agriculture. Les chaînes de montagnes qui entourent la vallée offrent des conditions idéales pour le développement de l'agriculture.

La situation géographique de la Vallée du Rhône est particulièrement intéressante. Elle se trouve à la confluence de plusieurs cours d'eau, ce qui contribue à l'irrigation et à l'agriculture. Les chaînes de montagnes qui entourent la vallée offrent des conditions idéales pour le développement de l'agriculture. Les chaînes de montagnes qui entourent la vallée offrent des conditions idéales pour le développement de l'agriculture.
Please join us!

Correct some errors in the old yearbooks of the SAC in Kokos

[Kollaboratives Korrektursystem]

http://kokos.cl.uzh.ch

Martin Volk   20.03.14