

Open Source-MÜ-Systeme

Seminar «Maschinelle Übersetzung» (HS 2011)

Simon Hafner & Hernani Marques

30. November 2011

Referatsübersicht

Problemstellung

Übersetzen von und nach Minderheitensprachen
Closed Source vs. Open Source Software

Herausforderungen & Lösungsansätze

Zwei OSS-MÜ-Systeme: Apertium und Moses

Apertium
Moses

Fragen

Minderheitensprachen

- ▶ Sprachen einer (kleinen) (Sub-)Kultur oder einer kleinen (geografischen) Region
- ▶ Beispiele: Klingonisch, Quechua, Baskisch, Esperanto

Problem: Bei vielen MÜ-Systemen keine Unterstützung

- ▶ Grosse/bekannte/kommerzielle MÜ-Systeme orientieren sich an Nachfrage/Rendite
- ▶ Datenmaterial für Minderheitensprachen ist nur spärlich vorhanden

Open Source Software (generell)



Abbildung: Open Source Software (Logos)¹

Lizenzen GPL, LGPL, CC, MIT/BSD/Apache, Shared Source u. a.
(libertär/viral, liberal, zweckbeschränkt)

Software Betriebssysteme, Browser, Spiele u. a.; auch:
MÜ-Systeme :)

¹URL: <http://cdn.lostintechology.com/wp-content/uploads/2010/12/opensource.jpg> (30.11.2011)

Closed Source Software (generell)

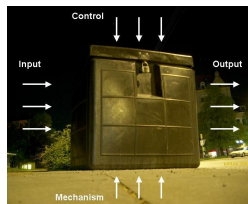


Abbildung: Closed Source Software (als Blackbox verstanden)²

Lizenzen EULA, Mehrplatzlizenzen, Shareware, Freeware u. a.

Software Betriebssysteme, Browser, Spiele u. a.; auch:
MÜ-Systeme :)

²URL: http://juid.de/images/Kaleidoskop/The_black_box.jpg (30.11.2011)

Open Source Software: Vorteile (1)

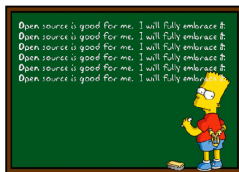


Abbildung: Bart: `*embraceOpenSource`³

- ▶ *Programmcode / Daten* sind kostenfrei (üblich)
- ▶ *Änderung/Erweiterung* des Programmverhaltens und der -funktionalität möglich
- ▶ *Nachvollzug* des Outputs möglich

³URL: <http://toomanytabs.com/blog/wp-content/uploads/2010/11/Bartopen.gif> (30.11.2011)

Open Source Software: Vorteile (2)

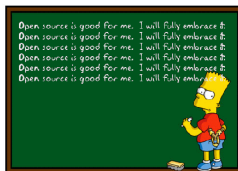


Abbildung: Bart: *embraceOpenSource*⁴

- ▶ *Kultur* des Teilens / kollektive Entwicklung
- ▶ *Partizipation* ist möglich und erwünscht

⁴URL: <http://toomanytabs.com/blog/wp-content/uploads/2010/11/Bartopen.gif> (30.11.2011)

Herausforderungen

- ▶ Minderheitensprache implementieren
- ▶ Kein kommerzieller Support
- ▶ Sparse-Data Problem
- ▶ stark agglutinierende Sprachen

Herausforderung (1)

Problem Kein kommerzieller Support

Lösung Auf den Schultern von Giganten

Lösung OSS Systeme

Herausforderung (2)

Problem Sparse-Data Problem

Problem stark agglutinierende Sprachen

Lösung Regeln implementieren

Lösung Zirkuläre Implikation [Forcada2006]

Apertium: Merkmale (1)



Abbildung: Apertium-Logo⁵

- ▶ Entwicklungsschwerpunkt in Spanien
- ▶ GPL-lizenziert

⁵URL: <http://www.eamt.org/corporate/logos/apertium.png> (30.11.2011)

Apertium: Merkmale (2)



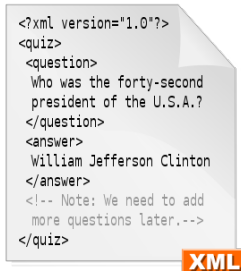
Abbildung: Apertium-Logo⁶

- ▶ RBMT
- ▶ Unterstützung für 28 Sprachpaare

⁶URL: <http://www.eamt.org/corporate/logos/apertium.png> (30.11.2011)

Technologien (Shallow-Transfer)

- ▶ Für syntaxähnliche Sprachen
- ▶ 1-Pass-Modell



```
<?xml version="1.0"?>
<quiz>
  <question>
    Who was the forty-second
    president of the U.S.A.?
  </question>
  <answer>
    William Jefferson Clinton
  </answer>
  <!-- Note: We need to add
    more questions later.-->
</quiz>
```

XML

Abbildung: Apertium-Module⁷

⁷URL:

Technologien (Advanced Transfer)

- ▶ Für syntaxunähnliche Sprachen
- ▶ 3-Pass-Modell: *Chunking* - *Interchunking* - *Postchunking*

Technologien (XML)

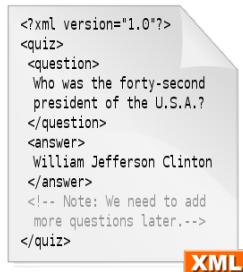


Abbildung: XML-Beispiel⁸

- ▶ Linguistische Daten werden in der XML gehalten
- ▶ Interoperabilität ist damit gesichert

⁸URL: <http://upload.wikimedia.org/wikipedia/commons/thumb/6/68/XML.svg/275px-XML.svg.png>

Phrase-Based Alignment



Abbildung: Phrases⁹

⁹URL: <http://www.statmt.org/moses/img/waiph5.png> (30.11.2011)

Phrase-Based Applied

Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap	by			green witch	
	no		slap		to the			
	did not give				to			
					the			
			slap			the witch		

Abbildung: Translations¹⁰

¹⁰URL: <http://www.statmt.org/moses/img/translation-options.png> (30.11.2011)

Beam Search

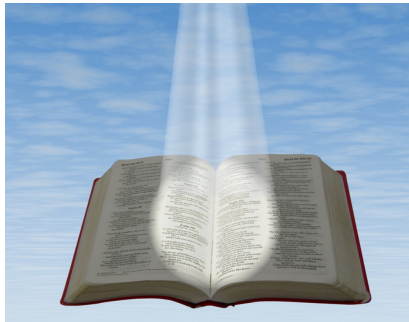


Abbildung: Beam Search¹¹

¹¹URL: <http://seo.mindandmouth.com/wp-content/uploads/2010/03/bible-illuminated-by-beam-of-light1.jpg?f8021c>
(30.11.2011)

Beam Example

Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap	by			green witch	
	no		slap		to the			
	did not give				to			
					the			
			slap			the witch		

Abbildung: Translations¹²

¹²URL: <http://www.statmt.org/moses/img/translation-options.png> (30.11.2011)

Beam Applied

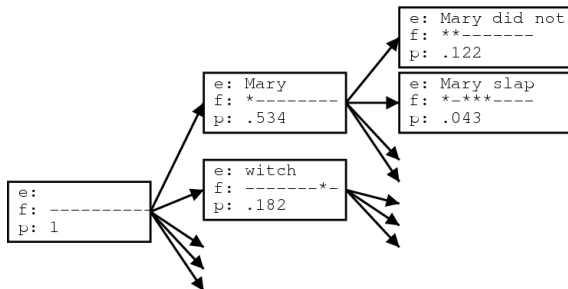


Abbildung: Translations¹³

¹³URL: <http://www.statmt.org/moses/img/beam.png> (30.11.2011)

Beam Search Data

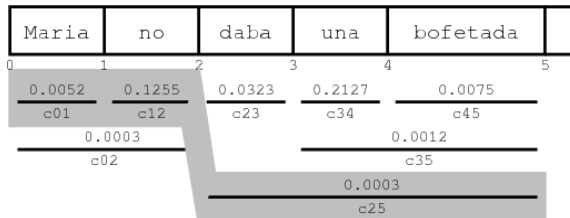


Abbildung: Future Cost¹⁴

¹⁴URL: <http://www.statmt.org/moses/img/phrase-beam-future-cost.png> (30.11.2011)

Factored

- ▶ Multi-Dimensionale Daten
- ▶ z.B. Morphologie

Factored Example

jllegh'egh

[SG][1PERS][NONE][VERB]legh[oneself]

Lektüreempfehlungen

- [Norvig1992] Norvig, Peter: *Paradigms of artificial intelligence programming: case studies in common LISP*. S.195-196. San Fransisco: Morgan.
- [Forcada2010] Forcada Mikel L. et al.: *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium*.
URL: <http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf> (30.11.2011)
- [Forcada2006] Forcada, Mikel L.: *Open-source machine translation: an opportunity for minor languages*. In: Strategies for developing machine translation for minority languages (5th SALT MIL workshop on Minority Languages).
URL: <http://www.dlsi.ua.es/~mlf/docum/forcada06p2.pdf> (30.11.2011)

Dalegh'a'¹⁶

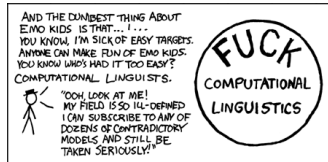


Abbildung: XKCD - Nr. 114¹⁵

¹⁵URL: <http://xkcd.com/114/> (30.11.2011)

¹⁶engl. "Do you see it?". URL: <http://en.wikibooks.org/wiki/Klingon/Grammar/Questions> (30.11.2011)