Vortrag SWP Gruppe 3

Translation Memories für automatische Patentübersetzung

Sabrina Mänz
Thomas Wangler

Institut für Computerlinguistik

22.07.2013

Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

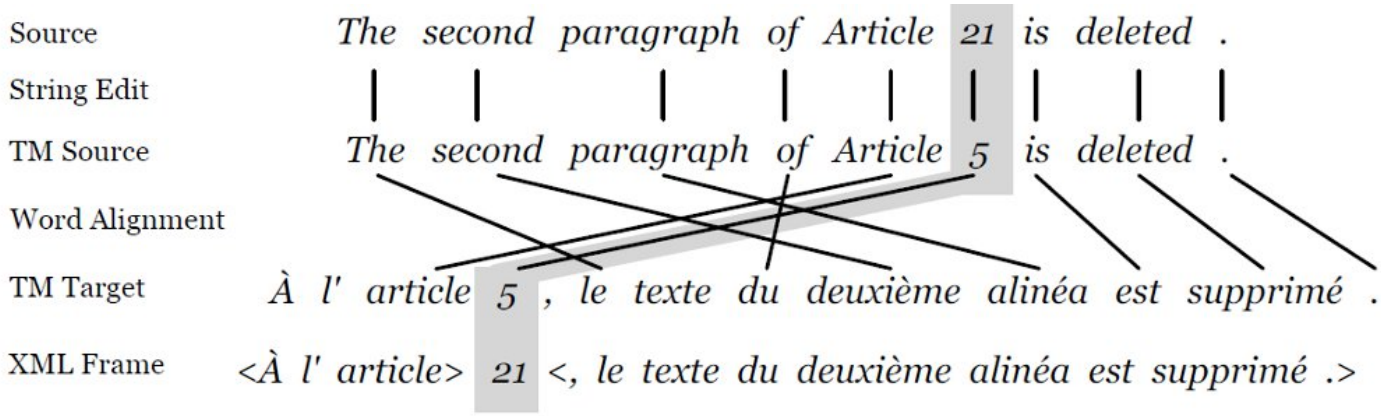
4 Ergebnisse

5 Resumee

6 Quellen

Erinnerung Aufgabenstellung

- Aus PatTR-Korpus SMT-System lernen
- Fuzzy Matches finden (TM)
- Bei geeignetem FMS merkliche Verbesserung gegenüber SMT-Baseline



Korpus

- Aufteilung anhand Metadaten (Filereader)
- **Testdaten**: Jahr 2000 - 2115 Sätze
- **Devsets**: 1996, 1998, 2002, 2004 - 2.037.564 Sätze
- **Trainingsset**: 5.271.065 Sätze
- pattr.claims insgesamt: 8.346.862 Sätze
- 1.036.118 Sätze erfüllen >100 Kriterium

Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

4 Ergebnisse

5 Resumee

6 Quellen

Vorbereitung des Korpus

- Entfernung von langen Sätzen: training/clean-corpus-n.perl
- tokenisieren: tokenizer/tokenizer.perl
- compound-splitter: generic/compound-splitter.perl
- lowercasen: tokenizer/lowercase.perl
- Satzgrenzen einfügen: irstlm/bin/add-start-end.sh
  Bsp.: <s>a device according to claim 8 characterised in that on the mounting or rear side the further disc-like cover (67) has a receiving opening (73) for receiveing the fixing element (68) which is in the form of a screw . </s >
- Vorformatieren für Hadoop und CreateXML
  Bsp.: 1|||<s>a device according to claim 8 [...] </s >

Baseline Model

- Build language Model: irstlm/bin/build-lm.sh
- Compile LM: irstlm/bin/compile-lm
- Build binarised LM: mosesdecoder/bin/build_binary
- Trainieren des Modells: training/train-model.perl
- moses.ini

Translating: vorrichtung nach anspruch 5 , dadurch gekennzeichnet , daß die weitere scheibenartige abdeckung (67) geringeren durchmessers auf der frontseite (64) [...]
BEST TRANSLATION: a device according to claim 5 , characterised in that the further disc-like cover (67) [...]

Testset I

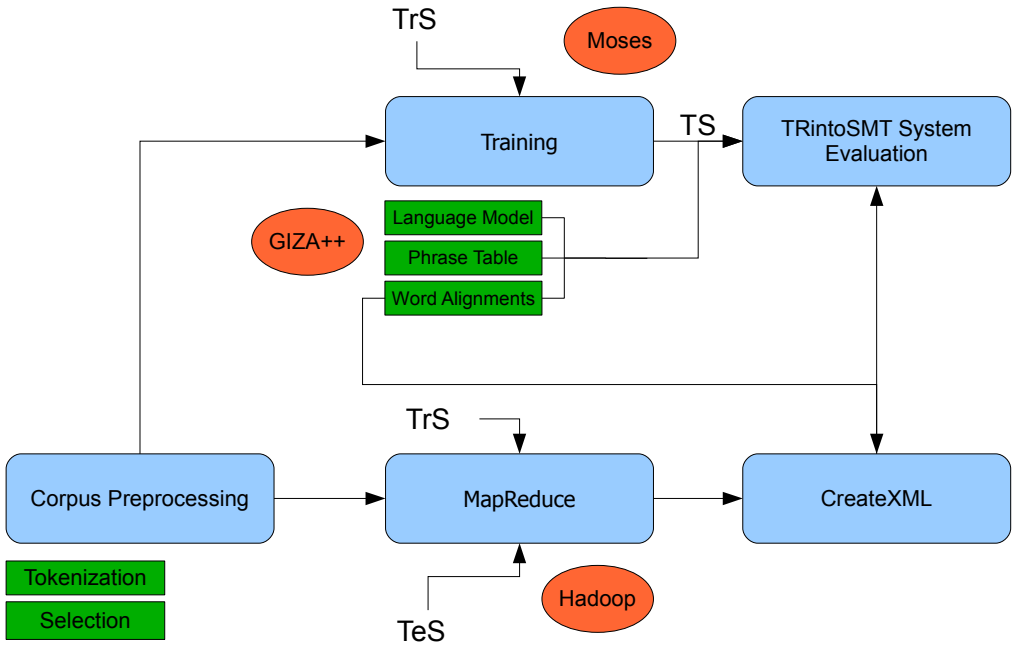
- Hadoop liefert Fuzzy Matches
  1001 3573116 0.85
  1009 4666468 0.714
- Keys aus hadoop-output-File extrahieren
- createFMString ausführen
- CreateXML ausführen
- Aufruf mosesdecoder: /mosesdecoder/bin/moses -xml-input exclusive -f ./moses.ini <xmlfile > translation.xml
- Englisches Corpus liefert Modellübersetzung
- Evaluation

Beispiel CreateXML

source: <s >verwendung nach anspruch 1 , worin die verbindung an ein polymer gebunden ist . </s >
tmSource: <s >verwendung nach anspruch 12 , wobei die verbindung freisetzbar an ein polymer gebunden ist . </s >
tmTarget: <s >the use according to claim 12 , wherein the compound is releasably bound to a polymer . </s >
aignments: 0-0 1-2 2-3 2-4 3-5 4-6 5-7 6-8 7-9 8-10 14-11 9-12 13-13 10-14 11-15 12-16 15-17 16-18
xml: <xml translation="use according to claim" >x </xml >1
<xml translation="," >x </xml >worin
<xml translation="the compound to a polymer bound is ." >
x </xml >

Implementierung

- kurze DEMO



Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

4 Ergebnisse

5 Resumee

6 Quellen

Evaluierung

- Rahmenbedingungen / Evaluierungstechnik:
  Vergleich von Sätzen, für die FM gefunden
- Vergleich von Ergebnissen von
  Baseline-SMT-Übersetzung vs. Korpus-Übersetzung und
  TMSystem-Übersetzung vs. Korpus-Übersetzung
- Evaluation:
  multi-bleu.perl -lc translation.modell < translation.xml
  multi-bleu.perl -lc translation.modell < translation.noxml

Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

4 Ergebnisse

5 Resumee

6 Quellen

Gefundene FM

- 312 Sätze aus Testset bei FMS >= 0,7 (15%)
- 1416 aus Testset bei FMS >= 0,5 (67%)

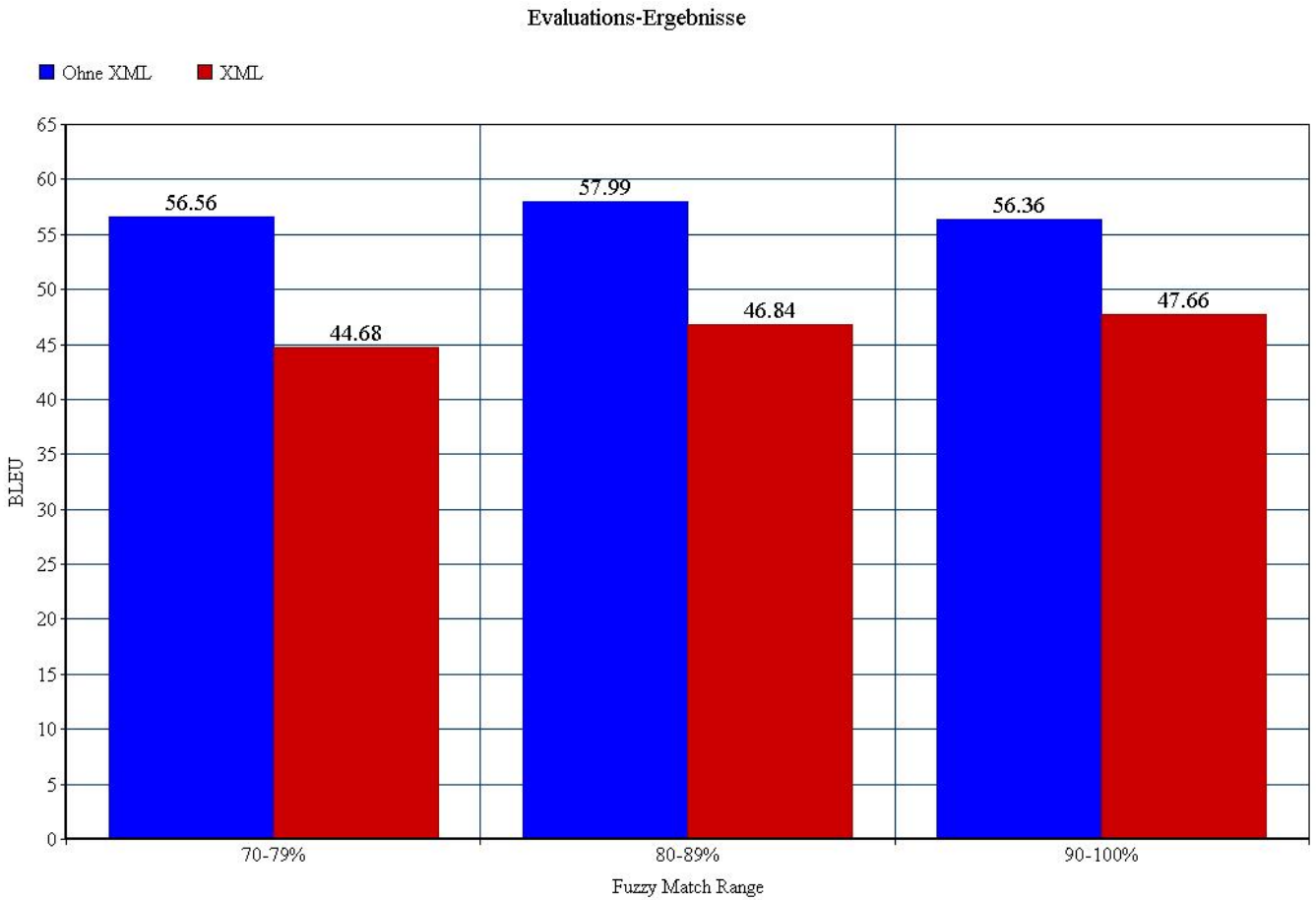
zur Erinnerung:
source: <s >verwendung nach anspruch 1 , worin die verbindung an ein polymer gebunden ist . </s >
tmSource: <s >verwendung nach anspruch 12 , wobei die verbindung freisetzbar an ein polymer gebunden ist . </s

Auswertung I

- Übersetzung **ohne xml**
  BLEU = 56.51
  (BP=0.972, ratio=0.973, hyp_ len=6500, ref_ len=6683)

- Übersetzung **mit xml**
  BLEU = 44.96
  (BP=0.953, ratio=0.955, hyp_ len=6379, ref_ len=6683)

Auswertung II



Auswertung zweites Set

- Korpusdaten (in Anlehnung an Acquis-Korpus, Koehn et al.):
  - Trainingsset: 1.165.867 Sätze
  - Testset: 4.107 Sätze
  - 63 FuzzyMatches (FMS >= 0,7)

- Übersetzung **ohne xml**
  BLEU = 53.81
  (BP=0.998, ratio=0.998, hyp_len=1258, ref_len=1261)
- Übersetzung **mit xml**
  BLEU = 41.86
  (BP=0.963, ratio=0.964, hyp_len=1215, ref_len=1261)

Auswertung Product - Koehn et al.

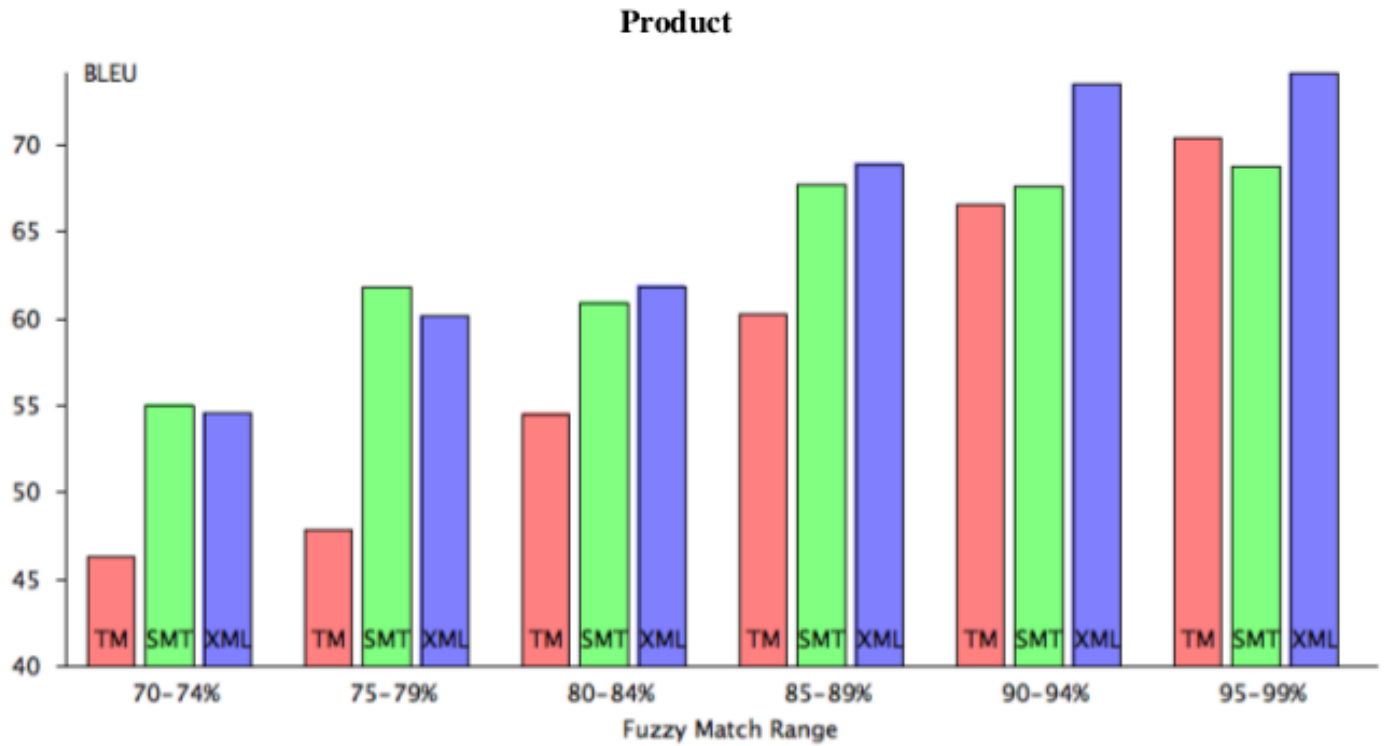


Figure 4: **Basic Results:** BLEU scores for different fuzzy match ranges. Our XML method performs best for sentence which have fuzzy matches of at least 80%, SMT is best below this threshold.

Mögliche Gründe

- Satzlänge im Testset (19.1 vs. 12 Wörter)
- Größe des Trainingssets
  - als Hauptgrund ausgeschlossen
- Paper: 83.461 Sätze (Product) vs. Projekt: 5.271.065 Sätze
- Struktur der xml-Konstruktion
- **Sprachpaar**: Englisch-Französisch vs. Deutsch-Englisch

Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

4 Ergebnisse

5 Resumee

6 Quellen

Herausforderungen I

- Projekt zu zweit

- Lauffähige Tools
- Moses-Parameter
- Dokumentation Hadoop

Herausforderungen II

- Implementierung des Pseudocodes

function **construct-xml:**

```
1:  xml = ""
2:  included = false
3:  start_t = −1
4:  for all target positions t ∈ [0; |t|[ do
5:      if !included AND matched-target(t) then
6:          start_t = −1
7:          included = true
8:      else if included AND (!matched-target(t) OR
        insertion[t]) then
9:          if start_t ≥ 0 then
10:             xml += "<xml translation="
11:             xml += t[start_t, t]
12:             xml += "> x </xml>"
13:         end if
14:         included = false
15:     end if
16:     xml += insertion[t]
17: end for
18: return xml
```

Gelungenes und Gelerntes

- Teamarbeit, google.code
- Technische Hürden überwunden
- Hadoop
- Notwendigkeit von Unittesting
- Dokumentation
- Devsets möglichst klein halten, Testen vor Verwendung
- Arbeiten mit großen Datenmengen kostet Zeit
- Zeitpläne helfen zur Orientierung
- Projekt läuft

Ausblick

- Optimierung wünschenswert
- Anwendung auf Gesamt-Pattr-Corpus
- Vergleich der verschiedenen Bereiche des Corpus - Unterschiede in der Nützlichkeit?
- Beide Richtungen übersetzen
- Anderer Ansatz für xmlConstruction
- z.B. Ordnen der tmTarget nach der source - Problem deletions

Inhaltsübersicht

1 Aufgabenstellung

2 Arbeitsschritte

3 Evaluierung

4 Ergebnisse

5 Resumee

6 Quellen

Quellen

- Philipp Koehn, Jean Senellart, 2011: Convergence of Translation Memory and Statistical Machine Translation, AMTA Workshop on MT Research and the Translation Industry.

- http://www.statmt.org/moses/?n=Moses.Baseline