Textdatenbanken

Sommersemester 2020

7. Vorlesung

- Paralleler Text -

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Aufgabenstellung

Es gibt für viele Texte Übersetzungen, in einigen günstigen Fällen ist alles frei verfügbar.

Frage: Wie lässt sich daraus ein zweisprachiges Wörterbuch konstruieren?

Oder: Wie lässt sich daraus das Wissen des Übersetzers „rekonstruieren“?

Arbeitsschritte:

Gegeben ist ein (möglicherweise großes) Dokument mit seiner Übersetzung.

Finde immer kleinere zusammengehörige Textteile:

- Absätze
- Sätze
- Wörter

Beispiel für Paralleltext: Vorher

English

According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved aboveaverage growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

French

Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures a celles de 1987, pour les boissons ~ base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. Uemploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le ler avril 1988 aprés une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Beispiel für Paralleltext: Nacher

Example from DE-News (8/1/1996)

English	German
Diverging opinions about planned tax reform	Unterschiedliche Meinungen zur geplanten Steuerreform
The discussion around the envisaged major tax reform continues .	Die Diskussion um die vorgesehene grosse Steuerreform dauert an .
The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .	Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der für 1999 geplanten Reform vorzuziehen .

Quellen für Paralleltext

Nicht:

- Bücher und ihre Übersetzung, weil i.d.R. nicht frei verfügbar.

Aber:

- Übersetzte und frei zugängliche Dokumente von Regierungen und internationalen Organisationen

Parallel Resources

- Newswire: DE-News (German-English), Hong-Kong News, Xinhua News (Chinese-English),
- Government: Canadian-Hansards (French-English), Europarl (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Spanish, Swedish), UN Treaties (Russian, English, Arabic, . . .)
- Manuals: PHP, KDE, OpenOffice (all from OPUS, many languages)
- Web pages: STRAND project (Philip Resnik)

Word-Level Alignments

- Given a parallel sentence pair we can link (align) words or phrases that are translations of each other:

Sentence Alignment

- If document D_e is translation of document D_f how do we find the translation for each sentence?
- The *n*-th sentence in D_e is not necessarily the translation of the *n*-th sentence in document D_f
- In addition to 1:1 alignments, there are also 1:0, 0:1, 1:n, and n:1 alignments
- Approximately 90% of the sentence alignments are 1:1

Sentence Alignment (c'ntd)

- There are several sentence alignment algorithms:
  - Align (Gale & Church): Aligns sentences based on their character length (shorter sentences tend to have shorter translations then longer sentences). Works astonishingly well
  - Char-align: (Church): Aligns based on shared character sequences. Works fine for similar languages or technical domains
  - K-Vec (Fung & Church): Induces a translation lexicon from the parallel texts based on the distribution of foreign-English word pairs.

Different Approached to Text Alignment

- ***Length-Based Approaches***: short sentences will be translated as short sentences and long sentences as long sentences.
- ***Offset Alignment by Signal Processing Techniques***: these approaches do not attempt to align beads of sentences but rather just to align position offsets in the two parallel texts.
- ***Lexical Methods:*** Use lexical information to align beads of sentences.

Length-Based Methods I: General Approach

- ***Goal:*** Find alignment A with highest probability given the two parallel texts S and T:

$$\mathit{argmax}_A\, P(A|S, T) = \mathit{argmax}_A\, P(A, S, T)$$

- To estimate the above probabilities, the aligned text is decomposed in a sequence of aligned beads where each bead is assumed to be independent of the others. Then

$$P(A, S, T) \approx \Pi_{k=1}^{K} P(B_k).$$

- The question, then, is how to estimate the probability of a certain type of alignment bead given the sentences in that bead.

Length-Based Methods II: Gale and Church, 1993

- The algorithm uses sentence length (measured in characters) to evaluate how likely an alignment of some number of sentences in L1 is with some number of sentences in L2.
- The algorithm uses a Dynamic Programming technique that allows the system to efficiently consider all possible alignments and find the minimum cost alignment.
- The method performs well (at least on related languages). It gets a 4% error rate. It works best on 1:1 alignments [only 2% error rate]. It has a high error rate on more difficult alignments.

Length-Based Methods II: Other Approaches

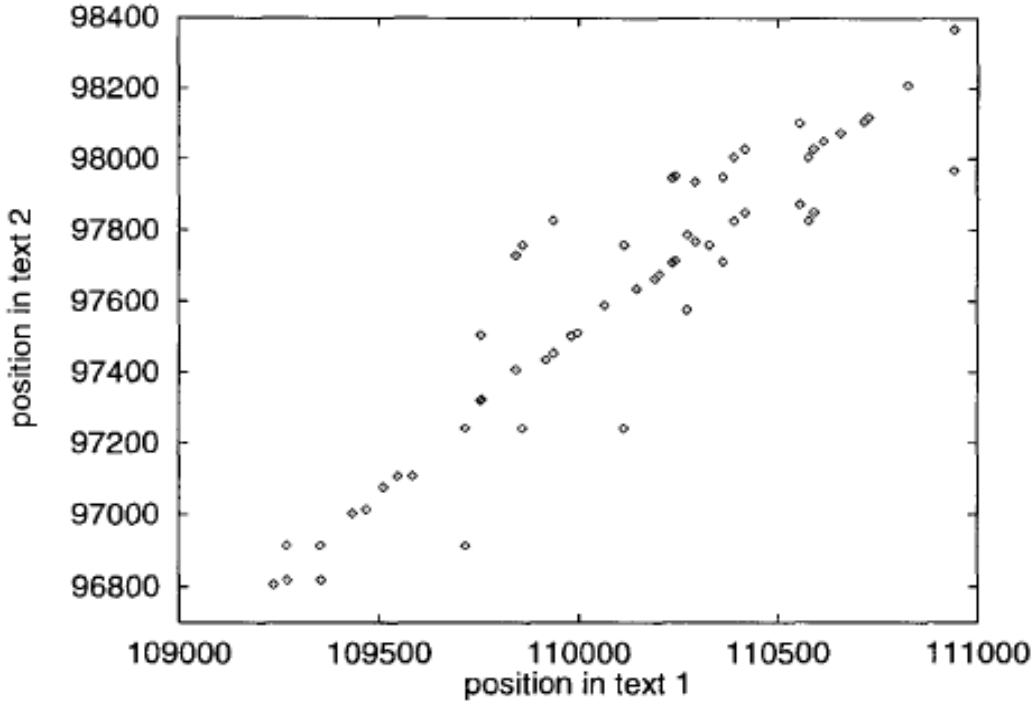
- ***Brown et al., 1991:*** Same approach as Gale and Church, except that sentence lengths are compared in terms of words rather than characters. Other difference in goal: Brown et al. Didn't want to align entire articles but just a subset of the corpus suitable for further research.
- ***Wu, 1994***: Wu applies Gale and Church's method to a corpus of parallel English and Cantonese Text. The results are not much worse than on related languages. To improve accuracy, Wu uses lexical cues.

Offset Alignment by Signal Processing Techniques I : Church, 1993

- Church argues that length-based methods work well on clean text but may break down in real-world situations (noisy OCR or unknown markup conventions)
- Church's method is to induce an alignment by using ***cognates*** (words that are similar across languages) at the level of character sequences.
- The method consists of building a dot-plot, i.e., the source and translated text are concatenated and then a square graph is made with this text on both axes. A dot is placed at (x,y) when there is a match. [Unit=4-grams].

Offset Alignment by Signal Processing Techniques II: Church, 1993 (Cont'd)

- Signal processing methods are then used to compress the resulting plot.
- The interesting part in a dot-plot is called the ***bitext maps***. These maps show the correspondence between the two languages.
- In the bitext maps, can be found faint, roughly straight diagonals corresponding to cognates.
- A heuristic search along this diagonal provides an alignment in terms of offsets in the two texts.



Offset Alignment by Signal Processing Techniques III: Fung & McKeown, 1994

- Fung and McKeown's algorithm works:
  - Without having found sentence boudaries.
  - In only roughly parallel text (with certain sections missing in one language)
  - With unrelated language pairs.
- The technique is to infer a small bilingual dictionary that will give points of alignment.
- For each word, a signal is produced, as an arrival vector of integer numbers giving the number of words between each occurrence of the word at hand.

Lexical Methods of Sentence Alignment I: Kay & Roscheisen, 1993

- Assume the first and last sentences of the texts align. These are the initial ***anchors***.
- Then, until most sentences are aligned:
  1. Form an ***envelope*** of possible alignments.
  2. Choose pairs of words that tend to co-occur in these potential partial alignments.
  3. Find pairs of source and target sentences which contain many possible lexical correspondences. The most reliable of these pairs are used to induce a set of ***partial alignments*** which will be part of the final result.

Lexical Methods of Sentence Alignment II: Chen, 1993

- Chen does sentence alignment by constructing a simple word-to-word translation model as he goes along.
- The best alignment is the one that maximizes the likelihood of generating the corpus given the translation model.
- This best alignment is found by using dynamic programming.

Lexical Methods of Sentence Alignment III: Haruno & Yamazaki, 1996

- Their method is a variant of Kay & Roscheisen (1993) with the following differences:
  - For structurally very different languages, function words impede alignment. They eliminate function words using a POS Tagger.
  - If trying to align short texts, there are not enough repeated words for reliable alignment using Kay & Roscheisen (1993). So they use an online dictionary to find matching word pairs

Word Alignment

- A common use of aligned texts is the derivation of bilingual dictionaries and terminology databases.
- This is usually done in two steps: First, the text alignment is extended to a word alignment. Then, some criterion, such as frequency is used to select aligned pairs for which there is enough evidence to include them in the bilingual dictionary.
- Using a χ^2 measure works well unless one word in L1 occurs with more than one word in L2. Then, it is useful to assume a one-to-one correspondence.
- Future work is likely to use existing bilingual dictionaries.

Cognates

Definitions of **cognates** on the Web:

- Words from two languages that are similar in spelling and meaning or sound and meaning
- Words that are similar in two or more languages as a result of common descent.
- Cognates are words from different languages which are related historically, eg English bath - German bad or English yoke - Hindi yoga. Beware FalseFriends however.

Extraction of cognates

- string comparison on the level of types in two parallel segments: Perl module String::Approx (Hietanainen 2002)
- high precision → cognates override bilingual lexicon

informatika	informatics
infrastrukture	infrastructure
instrumentacija	instrumentation
integracija	integrating
integrala	integral
iterativen	iterative
karakteristik	characteristics
kaskade	cascade
koeficient	coefficient
komponenta	component
koncentracijo	concentration
koncept	concept
konstanta	constant
konvergenca	convergence
koordinat	coordinates
linearne	linear
logisticxna	logistic
materiali	materials
matrika	Matrix

Soundex: Ähnlich klingende Wörter

Kodierungsschema für Buchstaben

1: B P F V
2: C S K G J Q X Z
3: D T
4: L
5: M N
6: R

Regeln:

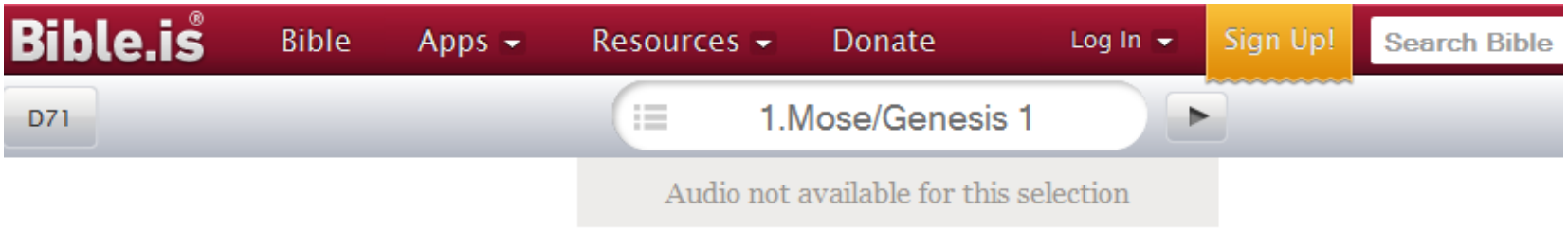
1. Ersten Buchstaben übernehmen
2. A E I O U W Y H löschen
3. Restliche Buchstaben entsprechend Kodierungsschema ersetzen
4. Dopplungen löschen
5. Auf 4 Zeichen kürzen, ggf. mit Nullen auffüllen.

Varianten: Reihenfolge 1,3,4,2,5
Schritt 5 weglassen (z.B. bei MySQL)

BibeIn

- Übersetzt in sehr viele Sprachen (>1000)
- In vielen davon im Netz verfügbar, oft bereits in utf8
- Quelle unter anderem: www.bible.is
  - Mehrere hundert Bibelversionen
  - Feste Ordnerstruktur für jede Sprache
  - /Gen/1 /Gen/2 …, daher leicht crawlbar

Bibeln



1 Im Anfang schuf Gott die Himmel (Im Hebr. steht das Wort "Himmel" immer in der Mehrzahl) und die Erde. **2** Und die Erde war wüst und leer, und Finsternis war über der (W. über der Fläche der) Tiefe; (Eig. eine rauschende, tiefe Wassermenge; so auch Kap. 7,11;8,2 2. Mo. 49,25) und der Geist Gottes schwebte über den Wassern. (W. über der Fläche der) **3** Und Gott sprach: Es werde Licht! und es ward Licht. **4** Und Gott sah das Licht, daß es gut war; und Gott schied das Licht von der Finsternis. **5** Und Gott nannte das Licht Tag, und die Finsternis nannte er Nacht. Und es ward Abend und es ward Morgen: erster Tag. (O. ein Tag) **6** Und Gott sprach: Es werde eine Ausdehnung inmitten der Wasser, und sie scheide die Wasser von den Wassern! **7** Und Gott machte die Ausdehnung und schied die Wasser, welche unterhalb der Ausdehnung, von den Wassern, die oberhalb der Ausdehnung sind. Und es ward also. **8** Und Gott nannte die Ausdehnung Himmel. Und es ward Abend und es ward Morgen: zweiter Tag. **9** Und Gott sprach: Es sammeln sich die Wasser unterhalb des Himmels an einen Ort, und es werde sichtbar das Trockene! Und es ward also. **10** Und Gott nannte das Trockene Erde, und die Sammlung der Wasser nannte er Meere. Und Gott sah, daß es gut war. **11** Und Gott sprach: Die Erde lasse Gras hervorsprossen, Kraut, das Samen hervorbringe, Fruchtbäume, die Frucht tragen nach ihrer Art, in welcher ihr Same sei auf der Erde! Und es ward also. **12** Und die Erde brachte Gras hervor, Kraut, das Samen hervorbringt nach seiner Art, und Bäume, die Frucht tragen, in welcher ihr Same ist nach ihrer Art. Und Gott sah, daß es gut war. **13** Und es ward Abend und es ward Morgen: dritter Tag. **14** Und Gott sprach: Es werden Lichter an der Ausdehnung des Himmels, um den

Bibeln

Eingeteilt in

- Bücher
- Kapitel
- Verse

Ein Vers ist oft ein Satz, manchmal auch mehr oder weniger.

Vorteil: Diese Nummerierung bleibt in den Übersetzungen erhalten.

Beispiel: **Einheitsübersetzung, Das Buch Genesis, Kapitel 19, Verse 17-19**

- **Gen 19,17:** Während er sie hinaus ins Freie führte, sagte er: Bring dich in Sicherheit, es geht um dein Leben. Sieh dich nicht um und bleib in der ganzen Gegend nicht stehen! Rette dich ins Gebirge, sonst wirst du auch weggerafft.
- **Gen 19,18:** Lot aber sagte zu ihnen: Nein, mein Herr,
- **Gen 19,19:** dein Knecht hat doch dein Wohlwollen gefunden. Du hast mir große Gunst erwiesen und mich am Leben gelassen. Ich kann aber nicht ins Gebirge fliehen, sonst lässt mich das Unglück nicht mehr los und ich muss sterben.

Parallele Verse

- Buch / Kapitel / Vers (und Name der Bibelübersetzung) in der Quelle.
- Beispiel: http://www.bible.is/GERD71/Gen/19#18

Beispiel in verschiedenen Sprachen:

Und Lot sprach zu ihnen: Nicht doch, Herr! http://www.bible.is/GERD71/Gen/19#18

En Lot zeide tot hen: Neen toch, Heere! http://www.bible.is/NLDDSV/Gen/19#18

Men Lot sagde til dem: “Ak nej, Herre! http://www.bible.is/DAND33/Gen/19#18

Но Лот сказал им: нет, Владыка! http://www.bible.is/RUSS76/Gen/19#18

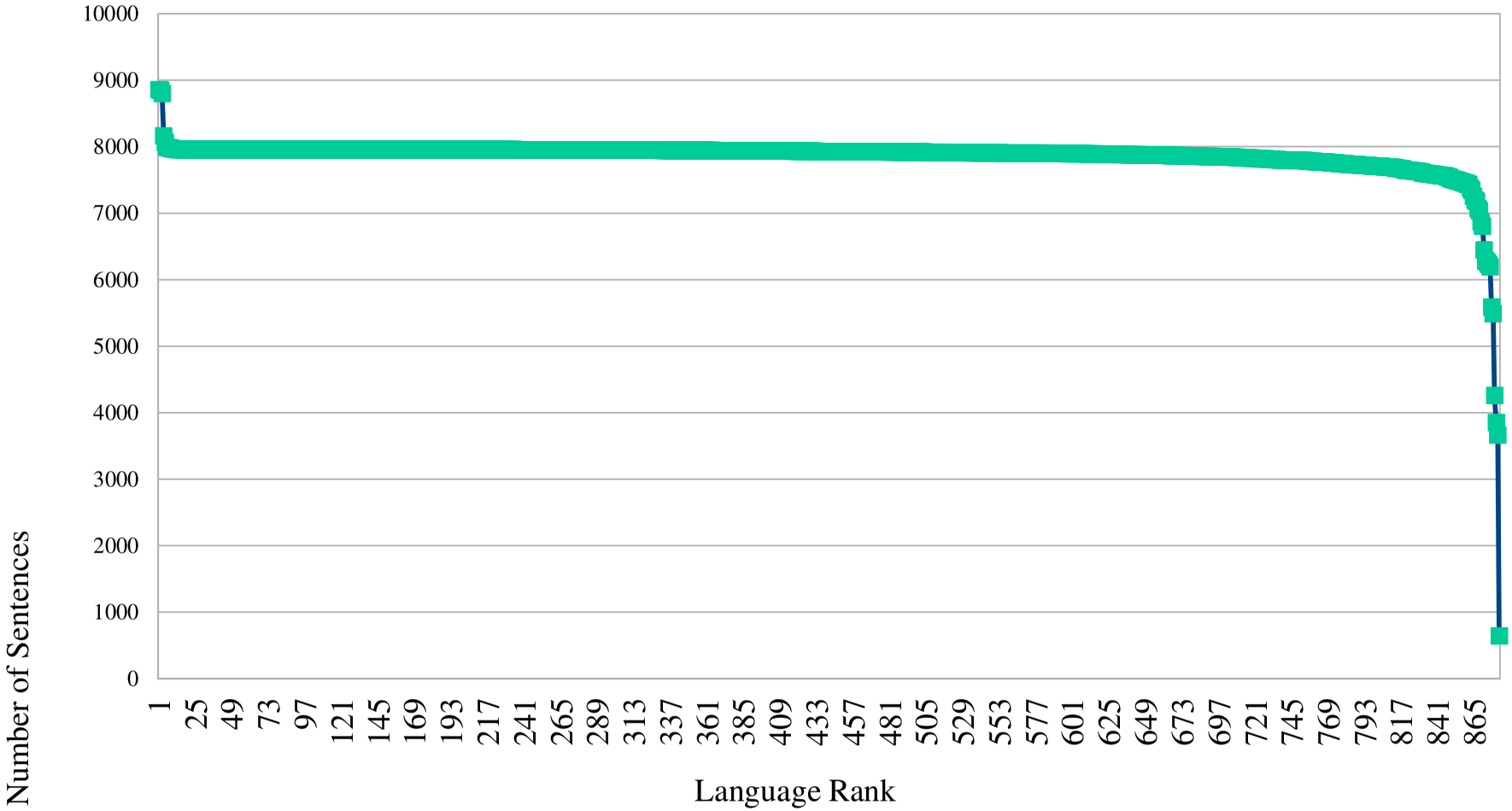
Tasol Lot i tok olsem, “Sori, Bikpela, mi no inap mekim olsem yu tok.
http://www.bible.is/TPIPNG/Gen/19#18

అయిత ఆ ఇదదరు మనుషయులత లతు ఇల చపపడు: “అయయలర, అంత దూరం పరుగతతమన ననను బలవంతం చయవదదు.
http://www.bible.is/TCWWTC/Gen/19#18

Korpora

- Verarbeitung:
  - Mit Blick auf Nutzung als Paralleltext:
    - Keine Sprachseparierung
    - Bibeln: Keine Zerlegung in Sätze
    - Kein musterbasiertes Entfernen von Sätzen

Korpora – Östen Dahl Bibeln



Bibeln - Vorteile

- Angemessene Menge Text (>7000 Sätze/Verse)
- Auch in Sprachen für die sonst keinerlei Ressourcen verfügbar sind
- Bereits der korrekten Sprache zugeordnet
- Paralleltext (versgenau)

Bibeln – vorhandene Ressourcen

- Bibeln:
  - 881 Versionen (Östen Dahl)
    - Meist ca. 8000 Verse,
  - 2000 Versionen aus verschiedenen Quellen: bible.is, papua-Bibeln
    - Deutlich mehr Text für große Sprachen
    - Kernbereich ebenfalls ca. 10k-7k Verse