Statistical Machine Translation LING-462/COSC-482 Week 2: Corpora sourcing, preparation and cleaning

Achim Ruopp

achim.ruopp@Georgetown.edu

Agenda

- Administrative items/Canvas
- Corpora sourcing, preparation and cleaning
  - Sourcing corpora
  - Building parallel corpora from aligned documents

- Break -

  - Crawling the web for parallel corpora
  - Corpus cleaning
  - Preparing corpora for MT training

What is a Parallel Text or Parallel Corpus?

- Translated text/documents in two languages
- Ideally sentence-aligned

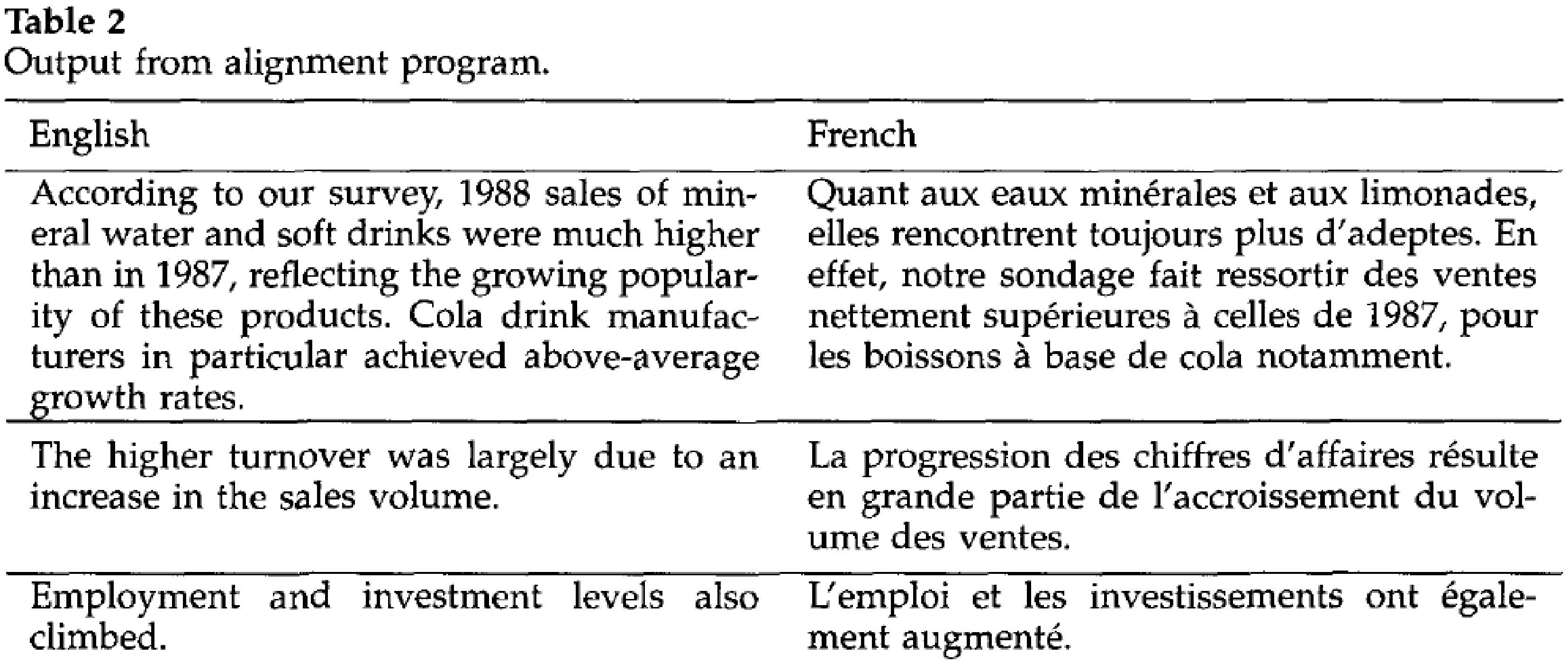


Table 2
Output from alignment program.

English	French
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 above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.

example from Gale & Church 1993

Parallel Corpus Formats
Plain Text

- Organization
  - Separate files for each language
  - Tab separated in a single file
  - Lines separated by line feed characters (Unix line endings) or line feed+carriage return characters (Windows line endings)
  - Sentence aligned with single sentences per line

Parallel Corpus Formats
Plain Text

- Unicode UTF-8 character encoding
  - Check for
    - Correct character encoding: Noël
    - Mojibake: NoÃ«l
  - Normalization Form Combined preferred
    - Noël (NFC) = Noe¨l (NFD)
  - No formatting markup or Unicode formatting characters
- Needs to be kept track of outside of files
  - What was the original source language?
  - Are these multi-lingual resources?
    - Makes pivoting possible

Text Editor

- Your favorite text editor
  - Can handle text files which are gigabytes large
  - Indicates file encoding
  - Indicates line endings
  - Allows conversion of encoding/line endings
- Options
  - vim Editor
  - Visual Studio Code
  - Not Notepad!

Text Bulk Processing Tools

- Character encoding detection
  - enca (Unix)
- Character conversion tool
  - iconv (Unix & Windows)
- Line ending conversion tool
  - tr, dos2unix
- Your favorite scripting language

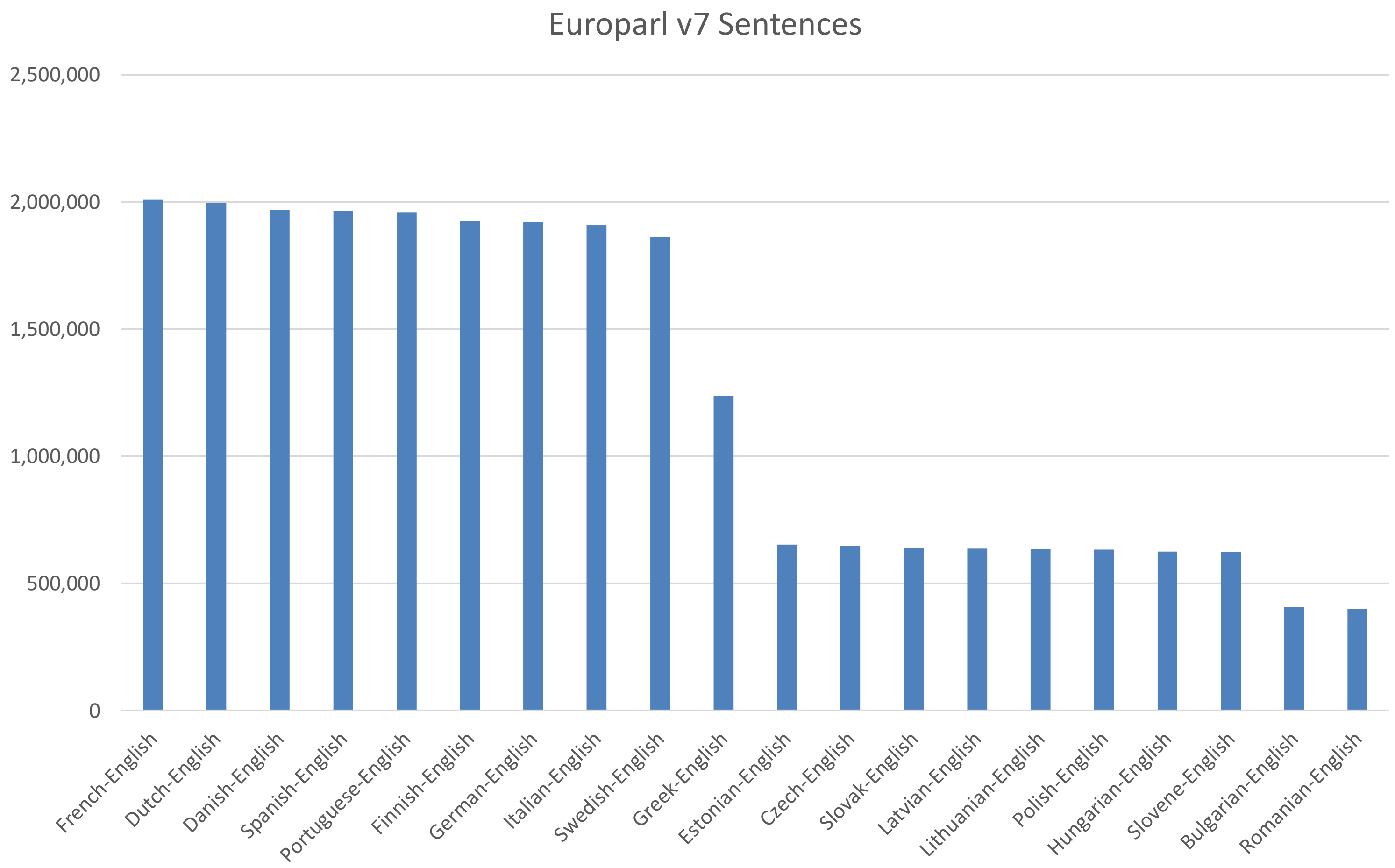
Parallel Corpus Formats
Translation Memory eXchange

- TMX v1.4b XML format (2005)
- Standard export format from computer aided translation tools
- Inherently keeps data and meta-data together
- Provides encoding/formatting control
- Supports inline markup for formatting/placeholders
- Virtually unsupported by MT toolkits
- Convert back and forth to plain text format
  - XML parser
  - Okapi Framework
  - …

Parallel Corpora - Governments

Name	Description	Domain	Aligned data (average)	Languages
Hansards	Canadian Parliament Proceedings	Legal/“General Domain”	1.3 million sentences	North American English, French
Europarl v7 (2012)	European Parliament Proceedings	Legal/“General Domain”	28 million source words	20 European languages to British English
EU language resources	Parallel corpora, terminology, Named Entities	Legal, Health, Culture, Education		24 European languages plus Turkish, Icelandic …
United Nations Parallel Corpus v1.0	Official records and other parliamentary documents	Legal	300 million+ words	English, French, Spanish, Russian, Chinese, Arabic

Europarl v7 Data Distribution



Parallel Corpora - Academic

Name	Description	Domain	Languages
WMT http://statmt.org/ wmtXX XX=06-18	Corpora from the Workshop/Conference on Machine Translation	News, biomedical, IT, …	Various with focus on European languages (because of funding)
OPUS	Free corpora collected by Jörg Tiedemann	IT, movie subtitles, medical	European, non-European for IT
Linguistic Data Consortium	Main repository for US-based corpus resources, some multilingual	News, others	English, Chinese, Arabic, …
European Language Resources Association	LDC-equivalent in Europe		Mostly European

Parallel Corpora - Industry

Name	Description	Domain	Languages
TAUS Data Cloud	Industry shared repository for translation memories (large part free for research use)	Several with slant to IT	All major languages
Translation Memories	Proprietary Translation Databases (mostly owned by translation buyer	Project-specific (great for incremental projects)	

Monolingual Corpus

- Needed to train target language model
- Use target language side of parallel corpus
- Better
  - Supplement with other target language data
    - Easy to obtain
    - Literature, technical manuals, etc.
    - Wikipedia
      - https://dumps.wikimedia.org/
      - https://radimrehurek.com/gensim/corpora/wikicorpus.html
    - Translation buyer documents in target language
  - Use to bias overall MT system to desired domain

BUILDING YOUR OWN PARALLEL CORPUS – FROM PARALLEL DOCUMENTS

Building a Parallel Corpus from Aligned Documents

1. Convert aligned documents to plain text
   - Caution: often deletes larger document structure
2. Segment sentences
3. Align sentences

- Assumes we have document alignment (e.g. based on document names)
- Full process automated with integrated tools
  - Open Source: LF Aligner
  - Commercial: Terminotix AlignFactory/YouAlign

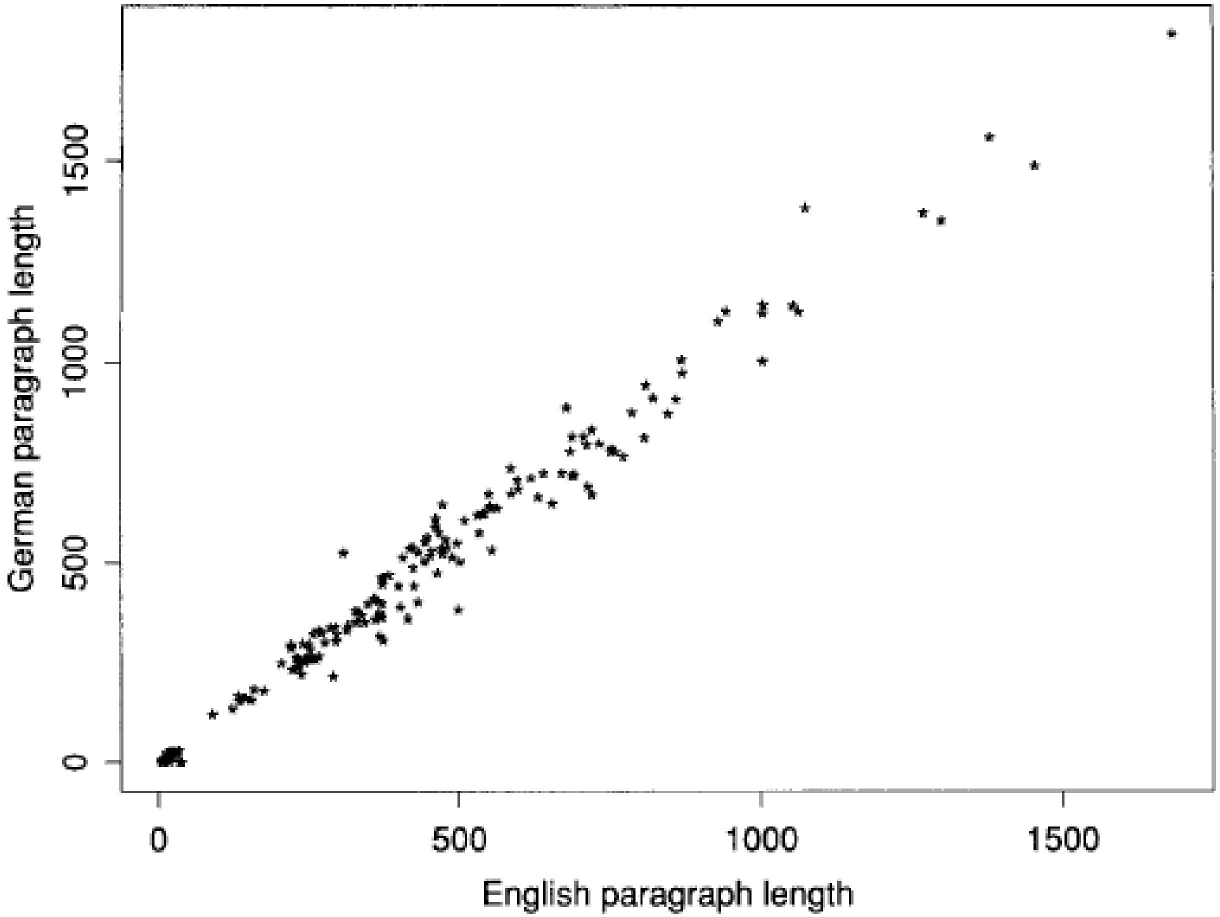
Building a Parallel Corpus Splitting Sentences

- Cannot just split on periods

  “Mr. Jones deposited $123.10 in his account.”

- Need list of abbreviations for language
  - Moses nonbreaking-prefixes for 25 languages and script split-sentences.perl
  - Other standard: Segmentation Rules eXchange (SRX)
    - More used in translation industry
    - Allows more granular segmentation rules with regular expressions
    - See paper “Using SRX Standard for Sentence Segmentation”

Building a Parallel Corpus
Automated Sentence Alignment



Gale & Church 1993: Paragraph length correlation in translated Union Bank of Switzerland (UBS) economic reports

Building a Parallel Corpus Automated Sentence Alignment

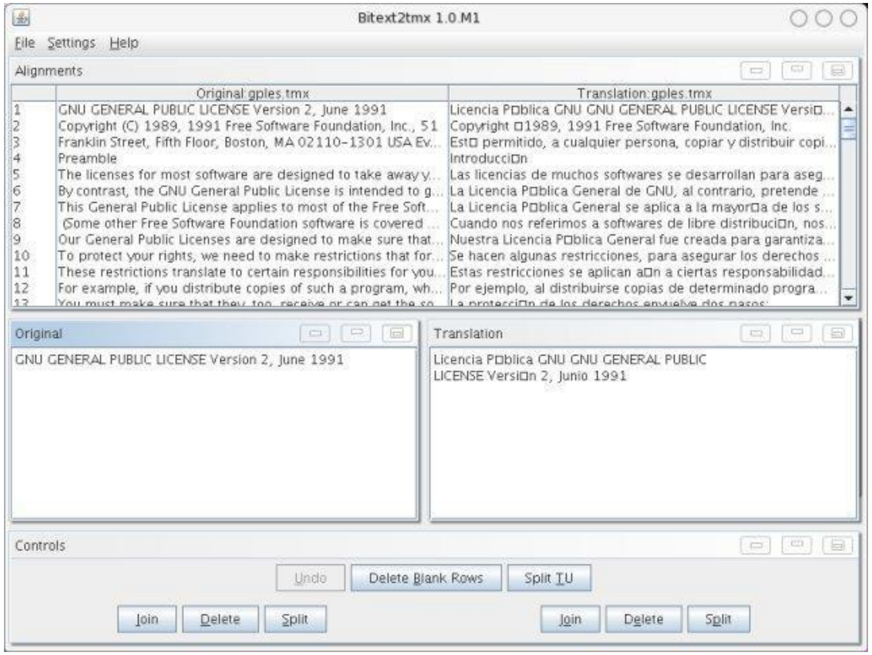
- Gale & Church 1993
  - probabilistic score is assigned to each proposed pair of sentences
    - Based on character counts of the two sentences
  - score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences
  - Language independent process
  - Language independent parameters
    - Not working for ideographic languages?

Building a Parallel Corpus Automated Sentence Alignment

- Gale & Church 1993 continued
  - Alignment scores
    - Allows filtering of low-confidence alignments
    - Often not made available for automatically aligned corpora
- Varga et. al. 2005: Hunalign
  - Uses dictionary when available
  - In absence of dictionary, falls back to sentence-length information, builds automatic dictionary and uses this in second pass

Building a Parallel Corpus Manual Sentence Alignment

- Manual correction step can follow automated alignment
- For alignment evaluation
- For machine translation evaluation sets
- Not for large training corpora
- Open source tools
  - Bitext2tmx
  - OmegaT



	Original:gples.tmx	Translation:gples.tmx
1	GNU GENERAL PUBLIC LICENSE Version 2, June 1991	Licencia P□blica GNU GNU GENERAL PUBLIC LICENSE Versi□...
2	Copyright (C) 1989, 1991 Free Software Foundation, Inc., 51	Copyright □1989, 1991 Free Software Foundation, Inc.
3	Franklin Street, Fifth Floor, Boston, MA 02110-1301 USA Ev...	Est□ permitido, a cualquier persona, copiar y distribuir copi...
4	Preamble	Introducci□n
5	The licenses for most software are designed to take away y...	Las licencias de muchos softwares se desarrollan para aseg...
6	By contrast, the GNU General Public License is intended to g...	La Licencia P□blica General de GNU, al contrario, pretende ...
7	This General Public License applies to most of the Free Soft...	La Licencia P□blica General se aplica a la mayor□a de los s...
8	(Some other Free Software Foundation software is covered ...	Cuando nos referimos a softwares de libre distribuci□n, nos...
9	Our General Public Licenses are designed to make sure that...	Nuestra Licencia P□blica General fue creada para garantiza...
10	To protect your rights, we need to make restrictions that for...	Se hacen algunas restricciones, para asegurar los derechos ...
11	These restrictions translate to certain responsibilities for you...	Estas restricciones se aplican a□n a ciertas responsabilidad...
12	For example, if you distribute copies of such a program, wh...	Por ejemplo, al distribuirse copias de determinado progra...
13	You must make sure that they, too, receive or can get the so	La protecci□n de los derechos envuelve dos pasos:

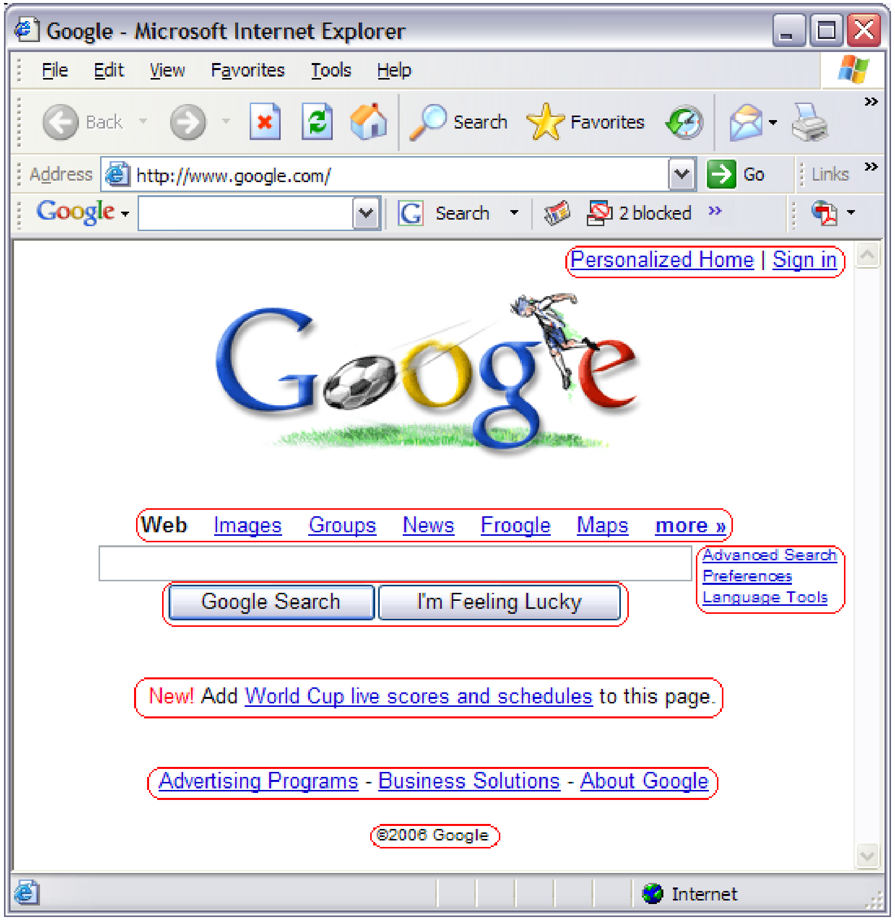
GNU GENERAL PUBLIC LICENSE Version 2, June 1991

Licencia P□blica GNU GNU GENERAL PUBLIC LICENSE Versi□n 2, Junio 1991

Image credit: Bitext2tmx

BUILDING YOUR OWN PARALLEL CORPUS – FROM THE WEB

Parallel Text on the Web


Google - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.google.com/ Go Links
Google Search 2 blocked
Internet

Personalized Home | Sign in

Web Images Groups News Froogle Maps more »

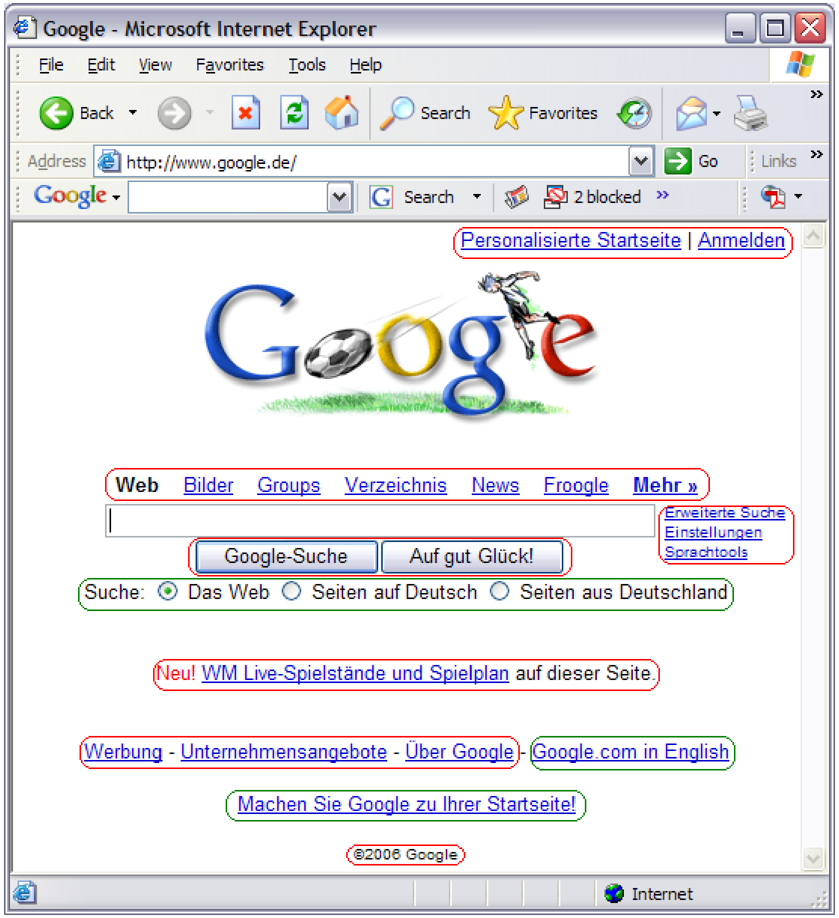
Advanced Search
Preferences
Language Tools

Google Search I'm Feeling Lucky

New! Add World Cup live scores and schedules to this page.

Advertising Programs - Business Solutions - About Google

©2006 Google


Google - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.google.de/ Go Links
Google Search 2 blocked
Internet

Personalisierte Startseite | Anmelden

Web Bilder Groups Verzeichnis News Froogle Mehr »

Erweiterte Suche
Einstellungen
Sprachtools

Google-Suche Auf gut Glück!

Suche: ● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Neu! WM Live-Spielstände und Spielplan auf dieser Seite.

Werbung - Unternehmensangebote - Über Google - Google.com in English

Machen Sie Google zu Ihrer Startseite!

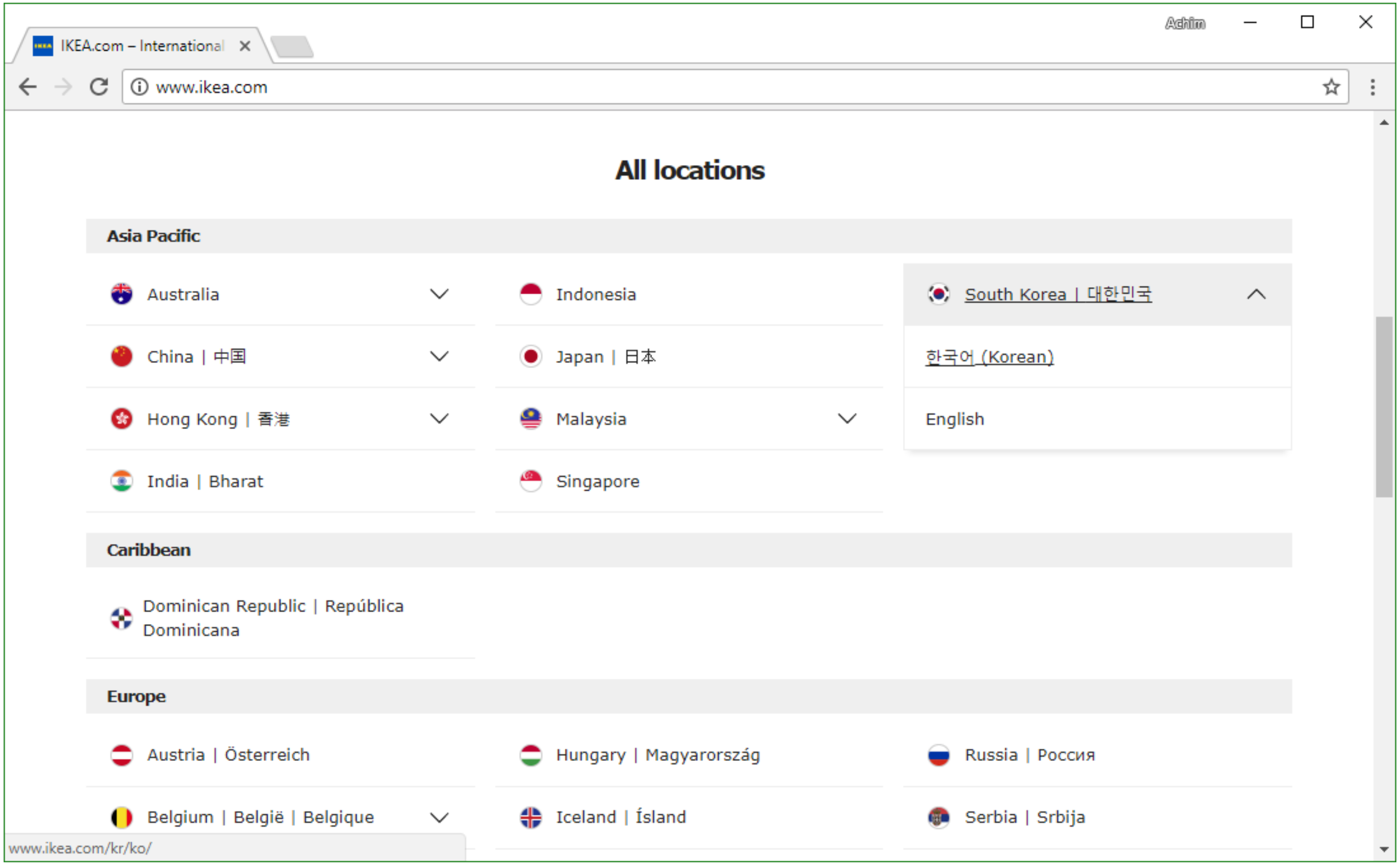
©2006 Google

Focused Crawling

- List of multilingual sites is known
- But document alignment on sites typically is not
  - Web servers/tools are not required to structure multil-lingual content in a certain way
- Resnik & Smith 2003
- Bitextor/Bicrawler
- ILSP Focused Crawler

Focused Crawling

Document Alignment with Link Structure



All locations

Asia Pacific

Australia	Indonesia	South Korea | 대한민국
China | 中国	Japan | 日本	한국어 (Korean)
Hong Kong | 香港	Malaysia	English
India | Bharat	Singapore	

Caribbean

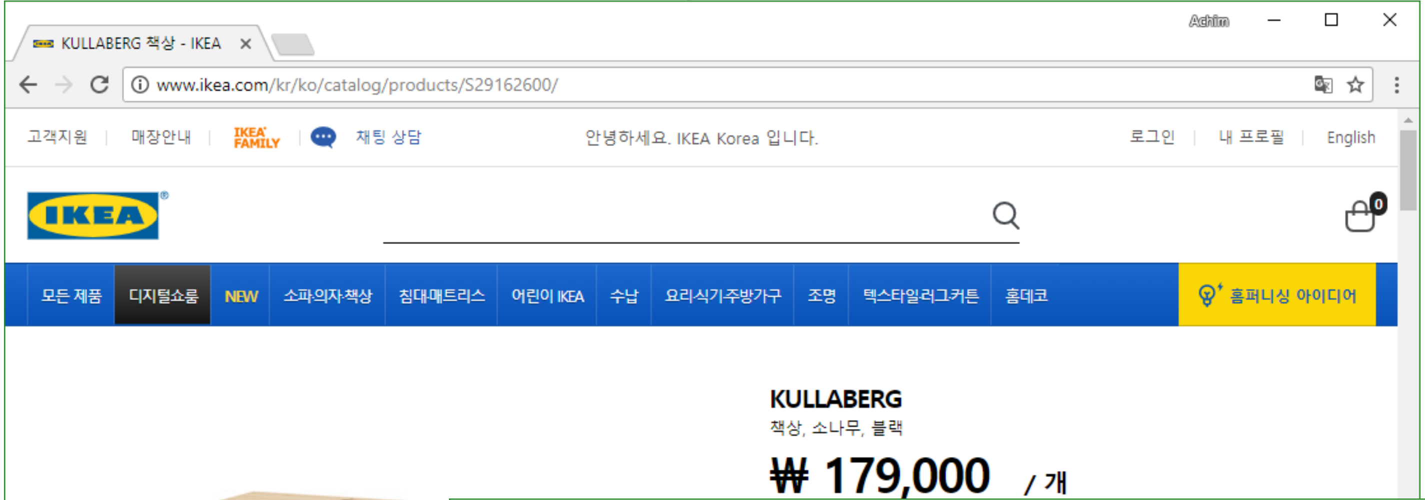
Dominican Republic | República Dominicana

Europe

Austria | Österreich	Hungary | Magyarország	Russia | Россия
Belgium | België | Belgique	Iceland | Ísland	Serbia | Srbija

Focused Crawling

Document Alignment with URL Patterns



KULLABERG 책상 - IKEA

www.ikea.com/kr/ko/catalog/products/S29162600/

고객지원 | 매장안내 | IKEA FAMILY | 채팅 상담

안녕하세요. IKEA Korea 입니다.

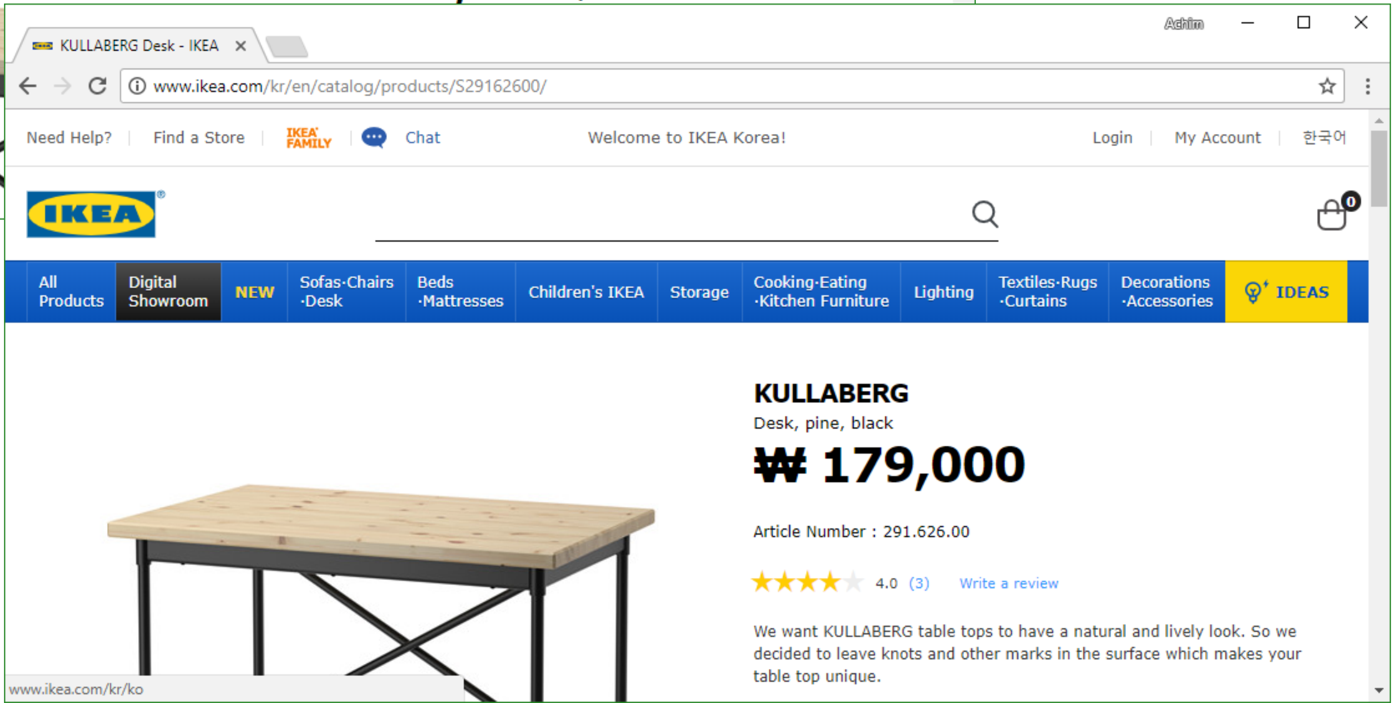
로그인 | 내 프로필 | English

모든 제품 | 디지털쇼룸 | NEW | 소파의자책상 | 침대매트리스 | 어린이 IKEA | 수납 | 요리식기주방가구 | 조명 | 텍스타일러그커튼 | 홈데코 | 홈퍼니싱 아이디어

KULLABERG

책상, 소나무, 블랙

₩ 179,000 / 개



KULLABERG Desk - IKEA

www.ikea.com/kr/en/catalog/products/S29162600/

Need Help? | Find a Store | IKEA FAMILY | Chat

Welcome to IKEA Korea!

Login | My Account | 한국어

All Products | Digital Showroom | NEW | Sofas·Chairs ·Desk | Beds ·Mattresses | Children's IKEA | Storage | Cooking·Eating ·Kitchen Furniture | Lighting | Textiles·Rugs ·Curtains | Decorations ·Accessories | IDEAS

KULLABERG

Desk, pine, black

₩ 179,000

Article Number : 291.626.00

★★★★☆ 4.0 (3) Write a review

We want KULLABERG table tops to have a natural and lively look. So we decided to leave knots and other marks in the surface which makes your table top unique.

www.ikea.com/kr/ko

- Standard BCP-47 language identifier for Korean in Korea is ko-KR
- No sibling pages here

Focused Crawling

Document Alignment with Content Similarity

- Simple things
  - same numbers, names or images in documents
  - often quite effective
- Use of dictionary
  - treat documents as bag of words
  - consider how many words in EN document have translations in FR document
- A bit more complex
  - semantic representations of documents content
  - bag of word vectors
  - neural network embeddings
- Major challenge: do this fast for n x m document pairs

Source: JHU Fall 2017 MT class © Philipp Koehn CC BY 3.0

Focused Crawling
Document Alignment with LCS

- Resnik & Smith 2013
- Serialize HTML structure of two candidate pages and calculate longest common subsequence

[START:HTML]	[START:HTML]
[START:TITLE]	[START:TITLE]
[Chunk:13]	[Chunk:15]
[END:TITLE]	[END:TITLE]
[START:BODY]	[START:BODY]
[START:H1]	
[Chunk:13]	
[END:H1]	
[Chunk:112]	[Chunk:122]

Focused Crawling
Document Alignment with LCS

- Calculate four values to determine alignment quality
  - *dp* The difference percentage, indicating nonshared material (i.e., alignmenttokens that are in one linearized file but not the other).
  - *n* The number of aligned nonmarkup text chunks of unequal length.
  - *r* The correlation of lengths of the aligned nonmarkup chunks.
  - *p* The significance level of the correlation *r*
- Can be combined with machine-learned decision tree

Focused Crawling Google's Content Matching

- Uszkoreit et. al. 2010: basic idea: translate everything into English, match large n-grams
- For each non-English document:
    1. Translate everything to English using MT
    2. Find distinctive ngrams
        a) rare, but not too rare (5-grams)
        b) used for matching only
- Build inverted index: ngram→documents
    - [cat sat on] → {[doc1, ES], [doc3, DE], ...}
    - [on the mat] → {[doc1, ES], [doc2, FR], ...}

Source: JHU Fall 2017 MT class © Philipp Koehn CC BY 3.0

Focused Crawling

- Once document pairs are found they can be aligned with the techniques described earlier
- Open source tools
  - Bitextor
  - Bicrawler
  - ILSP Focused Crawler
- WMT16 Bilingual Document Alignment Task

Broad Crawling

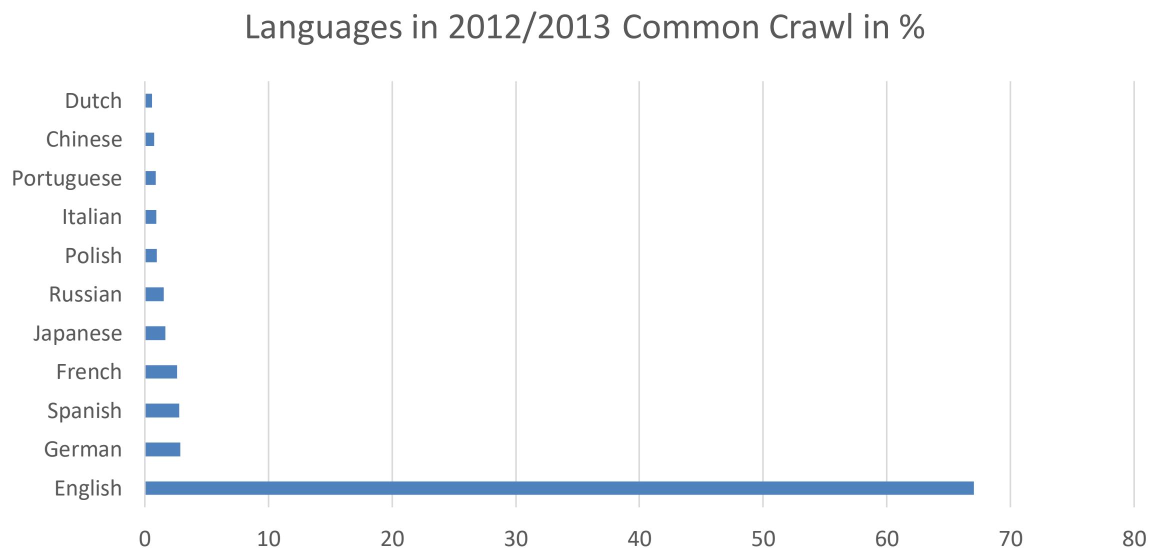
- Goal: Discovering web sites with parallel text on the entire web
- Could run own crawler
- Better: use Common Crawl
  - an open repository of web crawl data that can be accessed and analyzed by anyone
  - Monthly crawls of just below 3 billion web pages
  - Hosted on Amazon Web Services public datasets
  - Baseline: identify languages and match pages with URL patterns

Broad Crawling

- Goal: Discovering web sites with parallel text on the entire web
- Could run own crawler
- Better: use Common Crawl
  - an open repository of web crawl data that can be accessed and analyzed by anyone
  - Monthly crawls of just below 3 billion web pages
  - Hosted on Amazon Web Services public datasets
  - Baseline: identify languages and match pages with URL patterns

Broad Crawling
Common Crawl

- Monolingual data Buck, Heafield and van Ooyen (2014)



- Has turned much less English-centric since

Language Identification What language?

Muitas intervenções alertaram para o facto de a política dos sucessivos governos PS, PSD e CDS, com cortes no financiamento das instituições do Ensino Superior e com a progressiva desresponsabilização do Estado das suas funções, ter conduzido a uma realidade de destruição da qualidade do Ensino Superior público.

Source: JHU Fall 2017 MT class © Philipp Koehn CC BY 3.0

Language Identification
Character N-grams provide clues

Muitas intervenções alertaram para o facto de a política dos sucessivos governos PS, PSD e CDS, com cortes no financiamento das instituições do Ensino Superior e com a progressiva desresponsabilização do Estado das suas funções, ter conduzido a uma realidade de destruição da qualidade do Ensino Superior público.

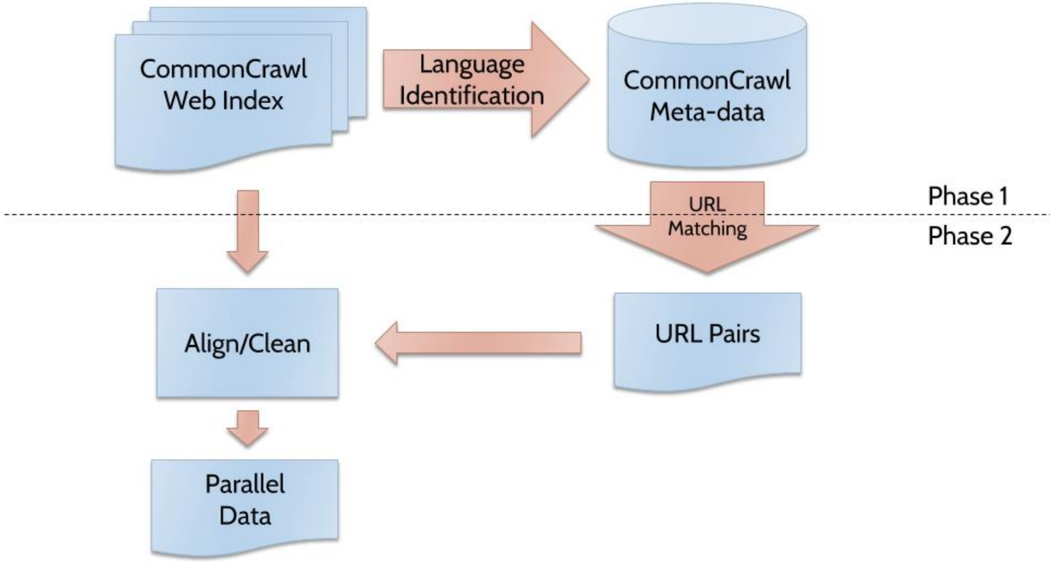
Source: JHU Fall 2017 MT class © Philipp Koehn CC BY 3.0

Language Identification Tools

- langid.py (Lui & Baldwin, ACL 2012)
  - 1-4 character grams, NaiveBayes, Feature Selection
- TextCat (based on Cavnar & Trenkle, 1994)
  - similar to langid.py
  - no Feature Selection
- Compact/Chromium Language Detector 2 (cld2; Google)
  - takes hints from tld, meta data
  - super fast
  - detects spans of text
  - Still takes about two weeks on 32 core machine to language tag one Common Crawl monthly crawl

Common Crawl
Matching URLs and Aligning Data

- ModernMT DataCollection



- Follow-on project: http://paracrawl.eu

CORPUS CLEANING

Parallel Corpus Cleaning
Language Independent

- Source and target segments identical
- Source segment empty
- Target segment empty
- Source and target segment lengths differ by more a certain percentage (e.g. 50%)
- Markup or placeholders do not match between source and target (if present)
- *Low sentence alignment score*

Parallel Corpus Cleaning Language-specific

- Source segment has incorrect language
- Target segment has incorrect language
- Spell checker reports large number of errors on source/target
- Target is machine translation of source
  - This can be hard to detect

Parallel Corpus Cleaning
Deduplication

- Deduplicate if source and target are duplicate
- Especially necessary for web data if boilerplate text has not been removed
- Overly aggressive?
  - For statistical systems duplicates could be significant for translation model
  - Maybe cut off beyond a certain absolute or relative repetition?

Parallel Corpus Cleaning Effectiveness

- Due to the large size and diverse origin of training corpora the overall quality of the training data can rarely be human evaluated
  - How should a representative sample be drawn?
- Rather
  - Test data cleaning in usage scenario by building multiple MT systems with different cleaned data versions
  - With baseline MT system

Parallel Corpus Cleaning Open Source Tool

- TMOP - Translation Memory Open-source Purifier
- Developed by Fondazione Bruno Kessler
- The goal of TMop is to identify and remove from the TM all the "bad" TUs, in which any of the two textual elements is either:
  1. syntactically poor,
  2. semantically different from the other,
  3. awkward according to some formatting criteria.

CORPUS PREPARATION

Corpus Preparation
Tokenization

- Separating
  - Words from each other
  - Punctuation from words

  “Mr. Jones deposited $123.10 in his account.” → “Mr. Jones deposited $123.10 in his account .”
- Fortunately we can use the same non-breaking prefixes/segmentation rules that we used for sentence segmentation
- East Asian lanuages
  - Chinese: Jieba
  - Japanese: ChaSen and MeCab

Corpus Preparation Casing

- Mostly lowercased data is used for MT training

  “Mr. Jones deposited $123.10 in his account .” → “mr. jones deposited $123.10 in his account .”

- Data can be recased with recasing statistical model
- Not lowercasing the data can lead to data sparseness
- Case matters (even more for languages like German)

  “Our neighborhood is getting a new Apple Store.”

  “Our neighborhood is getting a new apple store.”

Corpus Preparation Morphologically Rich Languages

- Decomposition into morphemes: smallest units in words that have semantic meaning
  - “unsolvable” → “un”-“solv”-“able”
- Addresses data sparsity for these languages
- Requires morphological analyzer
- Requires morphological composer
  - If morphologically rich language is target language
- Compound nouns: German

Corpus Preparation
Handling Named Entities

- Replacing named entities with placeholders

  "mr. jones deposited $123.10 in his account ." →

  "@@PERSON@@ depositied @@CURRAMT@@ in his account."

- Pro
  - Machine Translation Model need not be burdened with translation of items it is not suited for
- Con
  - Needs to be consistently applied across entire training corpus (can be hard if from different sources)
  - Loss of context information (e.g. reference of male pronoun)

Corpus Preparation
Handling Named Entities

- Needed when corpus needs to be anonymized
  - Privacy
- More often not applied to entire training corpus, but as decoder pre-/post-processing
- Literal Translation of entities
- Polyglot tool: NER for 40 languages

Assignments

- Language in 10 minutes
  - Starting in lecture 3 (1/25)
  - Proposal: 2 per lecture
  - Described in syllabus
- Homework 1: Quality of Machine Translation
  - To be published on Canvas
  - Time to complete: 2 weeks from when it is published

References

- Miłkowski M., Lipski J. (2011) Using SRX Standard for Sentence Segmentation. In: Vetulani Z. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Lecture Notes in Computer Science, vol 6562. Springer, Berlin, Heidelberg
- D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In Proceedings of the RANLP 2005, pages 590-596.
- Uszkoreit, Jakob & M. Ponte, Jay & Popat, Ashok & Dubiner, Moshe. (2010). Large Scale Parallel Document Mining for Machine Translation. 2. 1101-1109.
- Christian Buck and Kenneth Heafield and Bas van Ooyen, N-gram Counts and Language Models from the Common Crawl (2014), Proceedings of the Language Resources and Evaluation Conference, Reykjavik, Iceland