More Data Collection: Harvesting Parallel Documents from the Web

April 5, 2012

Thanks to Jakob Uszkoreit and Ashish Venugopal for many of today’s slides!
<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>فالفتعذيب لا يزال يمارس على نطاق واسع</td>
<td>Torture is still being practised on a wide scale.</td>
</tr>
<tr>
<td>وتتم عمليات الاعتقال والاحتجاز دون سبب بصورة روتينية</td>
<td>Arrest and detention without cause take place routinely.</td>
</tr>
<tr>
<td>وحان وقت التحلي بالبهجة والشجاعة السياسية</td>
<td>This is a time for vision and political courage</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>我国 能源 原材料 工业 生产 大幅度 增长.</td>
<td>China's energy and raw materials production up.</td>
</tr>
<tr>
<td>非国大 要求 阻止 更 多 被 拘留 人员 死亡.</td>
<td>ANC calls for steps to prevent deaths in police custody.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Goals for today’s lecture

• Understand how to mine bitexts from the web
• Web Crawling 101
• Review recent research into extracting parallel documents from the web and from unstructured collections
• What to do if you’re Google and you’re worried about harvesting your own machine translation output
The Web as a Parallel Corpus

• Old idea:


• Heuristically identify web pages that are potential translations of each other

• Download them

• Do filtering to check whether they are really translations
Heuristic identification

• Use link text
• If a page is written in English, and contains a link with the text Francais
• If the target page is written in French and contains a link with the text English
• Then the pair of documents may be translations of each other
Environmental Issues

Canadians are facing many issues that affect not only their health and well-being. Here are some resources to help you learn more about environmental issues in Canada, and to teach you how to take action.

Air
Climate Change
Habitat and Wildlife
Pollution and Waste
Water
Weather

Questions on the Environment

L'air
Changement climatique
Habitat et faune
Pollution et déchets
L'eau
La météo

Demandes d'accès à l'information complétées
Divulgation proactive

Changement climatique
Habitat et faune
Pyrénées (race caprine)

La chèvre des Pyrénées est une race caprine française originaire des Pyrénées. La Pyrénéenne est de taille moyenne : 75 à 85 cm au garrot pour un poids de 50 kg, et porte de longs poils, bruns ou noirs, parfois blancs. Elle peuple les Pyrénées depuis très longtemps et était autrefois associée aux troupeaux bovins et ovins, fournissant le lait aux bergers. Avec la modernisation de l'élevage, elle a failli disparaître dans la seconde moitié du XXe siècle. On s'intéresse toutefois de nouveau à elle depuis les années 1990, les effectifs remontent grâce au travail des conservatoires régionaux et, depuis 2004, de celui de l'association Chèvre de Race pyrénéenne en charge du programme de sauvegarde de la race.

On observe actuellement deux types d'élevage, les systèmes allaitants et les systèmes laitiers. Les premiers produisent des chevreaux bons à abattre, généralement à la période de Pâques, qui pèsent généralement autour de 15 kg. Les systèmes laitiers traient les chèvres à partir du sevrage précocement ou en hiver et se servent généralement de leur lait aux taux butyreux et protéiques corrects pour fabriquer du fromage, crottin ou tomme des Pyrénées. Les chevreaux ne sont pas très bien conformés et la production de lait par chèvre reste bien en deçà de celle des races spécialisées. Toutefois, la chèvre des Pyrénées a l'avantage d'être très rustique et de pouvoir valoriser une végétation médicée, dans des conditions climatiques parfois très rude. Elle permet de maintenir certains paysages ouverts en empêchant qu'ils ne s'embroussaient.
Check for translation equivalence

• How would you check to see if two documents were translations of each other or not?

• How would your strategy differ if
  – you didn’t have any bilingual resources
  – you had a normal bilingual dictionary
  – you had a small amount of bitexts already

• Discuss with your neighbor
Page structure similarity

<HTML>
<TITLE>Emergency Exit</TITLE>
<BODY>
<H1>Emergency Exit</H1>
If seated at an exit and

<TITLE>Sortie de Secours</TITLE>
<BODY>
Si vous êtes assis à côté d’une ...

The aligned linearized sequence would be as follows:

[START:HTML]     [START:HTML]
[START:TITLE]     [START:TITLE]
[Chunk:13]        [Chunk:15]
[END:TITLE]       [END:TITLE]
[START:BODY]      [START:BODY]
[START:H1]        [START:H1]
[Chunk:13]        [Chunk:15]
[END:H1]          [END:H1]
[Chunk:112]       [Chunk:122]
• % of non-shared material
• number of aligned non-markup text chunks that are different in length
• correlation of lengths of the text chunks
• significance level of the correlation

— Set the value of each of those elements empirically against a set of manually classified real-world pages
Bilingual dictionary

• Use a bilingual dictionary to do a word-for-word lookup of all the words in document A, compare them to document B

\[
similarity(A, B) = \frac{\text{number of translation token pairs}}{\text{number of tokens in A}}
\]

• In addition to dictionary translations, can also count identical strings (numbers and names) or near identical strings (cognates)
What about translated URLs?

www.banqueducanada.ca/2012/04/discours/vieillir-en-beaute-inevitable-evolution/
www.bankofcanada.ca/2012/04/speeches/aging-gracefully-canadas-inevitable/
<table>
<thead>
<tr>
<th>Site Address</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>rparticle.web-p.cisti.nrc.ca</td>
<td>93236</td>
</tr>
<tr>
<td><a href="http://www.ec.gc.ca">www.ec.gc.ca</a></td>
<td>53973</td>
</tr>
<tr>
<td><a href="http://www.hc-sc.gc.ca">www.hc-sc.gc.ca</a></td>
<td>52318</td>
</tr>
<tr>
<td>portal.unesco.org</td>
<td>45118</td>
</tr>
<tr>
<td><a href="http://www.cra-arc.gc.ca">www.cra-arc.gc.ca</a></td>
<td>42737</td>
</tr>
<tr>
<td><a href="http://www.dfo-mpo.gc.ca">www.dfo-mpo.gc.ca</a></td>
<td>34617</td>
</tr>
<tr>
<td><a href="http://www.canadianheritage.gc.ca">www.canadianheritage.gc.ca</a></td>
<td>29445</td>
</tr>
<tr>
<td><a href="http://www.idrc.ca">www.idrc.ca</a></td>
<td>28170</td>
</tr>
<tr>
<td><a href="http://www.agr.gc.ca">www.agr.gc.ca</a></td>
<td>26823</td>
</tr>
<tr>
<td><a href="http://www.dfait-maeci.gc.ca">www.dfait-maeci.gc.ca</a></td>
<td>21255</td>
</tr>
<tr>
<td><a href="http://www.forces.gc.ca">www.forces.gc.ca</a></td>
<td>19827</td>
</tr>
<tr>
<td><a href="http://www.ic.gc.ca">www.ic.gc.ca</a></td>
<td>16922</td>
</tr>
<tr>
<td><a href="http://www.ceaa-acee.gc.ca">www.ceaa-acee.gc.ca</a></td>
<td>16492</td>
</tr>
<tr>
<td><a href="http://www.gg.ca">www.gg.ca</a></td>
<td>16289</td>
</tr>
<tr>
<td><a href="http://www.canadianencyclopedia.ca">www.canadianencyclopedia.ca</a></td>
<td>15002</td>
</tr>
<tr>
<td>www2.parl.gc.ca</td>
<td>14380</td>
</tr>
<tr>
<td><a href="http://www.fin.gc.ca">www.fin.gc.ca</a></td>
<td>14089</td>
</tr>
<tr>
<td><a href="http://www.aecb.org">www.aecb.org</a></td>
<td>13706</td>
</tr>
<tr>
<td><a href="http://www.cihr-irsc.gc.ca">www.cihr-irsc.gc.ca</a></td>
<td>13264</td>
</tr>
<tr>
<td><a href="http://www.cprn.org">www.cprn.org</a></td>
<td>12161</td>
</tr>
<tr>
<td><a href="http://www.civilisations.ca">www.civilisations.ca</a></td>
<td>12145</td>
</tr>
<tr>
<td><a href="http://www.cbsa.gc.ca">www.cbsa.gc.ca</a></td>
<td>11632</td>
</tr>
<tr>
<td><a href="http://www.cbsa-asfc.gc.ca">www.cbsa-asfc.gc.ca</a></td>
<td>11632</td>
</tr>
<tr>
<td><a href="http://www.hockeycanada.ca">www.hockeycanada.ca</a></td>
<td>11005</td>
</tr>
<tr>
<td><a href="http://www.crr.ca">www.crr.ca</a></td>
<td>10382</td>
</tr>
<tr>
<td><a href="http://www.commonlaw.uottawa.ca">www.commonlaw.uottawa.ca</a></td>
<td>10338</td>
</tr>
<tr>
<td><a href="http://www.ourroots.ca">www.ourroots.ca</a></td>
<td>10150</td>
</tr>
<tr>
<td><a href="http://www.cws-scf.ec.gc.ca">www.cws-scf.ec.gc.ca</a></td>
<td>9224</td>
</tr>
<tr>
<td><a href="http://www.elections.ca">www.elections.ca</a></td>
<td>8440</td>
</tr>
<tr>
<td><a href="http://www.collectionscanada.ca">www.collectionscanada.ca</a></td>
<td>8099</td>
</tr>
</tbody>
</table>
Web Crawling 101

- Mirror web sites
- Extract text page contents
- Perform language ID
- Segment into sentences
- Align document pairs
- Align sentences
- Remove duplicates
Mirror web sites

• We would like to crawl the web, saving pages to extract translated documents from
• Useful cross-platform GNU utility called wget
• Basic usage to download a single file:
  
  wget http://europa.eu/

• Download an entire web site, preserving directory structures:
  
  wget --mirror http://europa.eu/
There is a protocol that web sites use to instruct search engines and other web crawlers not to index certain pages.

Sites contain a file called robots.txt that indicates who is allowed to look at what.
That’s robo-prejudice!

- wget lets you ignore this protocol:
  ```bash
  wget -robots=off --mirror http://akhbarlive.com/
  ```
- Some sites will block wget directly, you can pretend to be some other browser:
  ```bash
  wget -robots=off --mirror -U "Mozilla/5.0 (compatible; Konqueror/3.2; Linux)"
  http://akhbarlive.com
  ```
- Don’t do this. But if you do, please do this too:
  ```bash
  wget --wait=5 --random-wait --limit-rate=512k --timeout=5
  -robots=off --mirror -U "Mozilla/5.0 (compatible; Konqueror/3.2; Linux)"
  http://akhbarlive.com
  ```
• For bilingual parallel corpora, we really only care about the text. HTML markup will mess us up.
• Convert web pages to text (surprisingly not easy)
• I use two programs
  – Apple’s textutil for HTML and Word
  – XPDF for PDF
• How do we know that a page is written in the language that we are expecting?

• HTML “meta” tag with ISO 639 2-letter language codes:

  `<meta http-equiv="content-language" content="en">`
  `<meta http-equiv="content-language" content="fr">`

• This meta-data is often missing or in accurate

• Statistical NLP to the rescue!
• Intuition: some character strings are more probable in one language than in others

<table>
<thead>
<tr>
<th>Language</th>
<th>char sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>vnd</td>
</tr>
<tr>
<td>English</td>
<td>ery</td>
</tr>
<tr>
<td>French</td>
<td>eux</td>
</tr>
<tr>
<td>Gaelic</td>
<td>mh</td>
</tr>
<tr>
<td>German</td>
<td>der</td>
</tr>
<tr>
<td>Italian</td>
<td>cchi</td>
</tr>
<tr>
<td>Portuguese</td>
<td>seu</td>
</tr>
<tr>
<td>Serbo-croat</td>
<td>lj</td>
</tr>
<tr>
<td>Spanish</td>
<td>ir</td>
</tr>
</tbody>
</table>
Dunning (1994)

\[
p(S \mid A) = p(s_1 \ldots s_k \mid A) \prod_{i=k+1}^N p(s_i \mid s_{i-k} \ldots s_k \mid A)
\]
• But Prof. Callison-Burch, Yahoo! answers.com tells me that this is a 99.66% of the time this is super easy to do...
Sentence segmenters

• NLTK has one called PUNKT that is trainable to other languages

• Download several from the WMT workshops
  – http://statmt.org/wmt08/scripts.tgz
• Write a regular expression to find pairs of URLs that are equivalent (s/_e/_f/) and see if there are matching files from your crawl
• Use link structure across pages with the STRAND trick
• Validate that the document pairs are plausible
Align sentences

• After we have identified parallel documents we need to align the sentences within them
• This is not straightforward because human translators do not always translate things in a 1-to-1 fashion
  – Sentences tend to be translated in same order linear
  – Can join two sentences into one
  – Can split one sentence into two
  – Can omit a sentence (by mistake)
  – Can add a sentence (for elaboration)
Sentence alignment

• Use dynamic programming to find the best alignment between sentences in a document
  – Use sentence lengths in absence of other info
  – Use bilingual dictionaries to score alignments
  – Use Model-1 probabilities to score alignments
• Jason Smith will discuss this topic in more depth on Tuesday
• Open source tool from Bob Moore:
  http://research.microsoft.com/en-us/downloads/aaf5d5d5-4d5c-49b2-8a22-f7055113e656/
Remove duplicates

• With large scale crawls, there are often duplicates at page level or sub-page level
  – with www. prefix and without
  – printable versions of articles and regular versions
  – template text like budgets that vary only in $ amount
  – navigation gets replicated across an entire site
  – remove text that is left untranslated

• We would like to remove duplicate pages, or better yet, duplicate sentences

• Problem: too much data to store in a HashTable/HashSet and check strings against
The birdseye figured grain in sugar maple (Acer saccharum). Literature review, nomenclature, and structural characteristics

Don C. Bragg


ABSTRACT

Little is known about the "birdseye" figured grain of sugar maple (Acer saccharum Marsh.). This paper clarifies and expands the discussion of birdseye sugar maple by describing the similarities and differences with figured grains in other species, as well as discussing important features of its peculiar anatomy. Sections are also provided that discuss the proposed causes of the birdseye grain, detail birdseye sugar maple's geographic distribution, and address what is known about genetics and birdseye maple. Possible variations on the birdseye theme (e.g., roundeye, bongeal, cat's paw, distorted) are documented, and a new set of descriptive terminology is established. Finally, further observations and speculations on the birdseye phenomena are provided, and research directions are suggested.

Cited by
View all 2 citing articles

RéSUMÉ

On connaît peu de chose à propos du grain de l'érable à sucre (Acer saccharum) présent dans les mouchetures. Cet article clarifie et élargit la discussion au sujet des similitudes et des différences avec le grain texturé chez d'autres espèces d'érables, ainsi que dans les différentes parties de son anatomie particulière. Des sections sont également incluses qui discutent des causes proposées du grain de l'érable à sucre, décrivent la distribution géographique de l'érable à sucre, et abordent ce que l'on sait sur la génétique et l'érable à sucre. Les variations possibles des mouchetures texturées (p. ex., rondes, déformées) sont présentées, et une nouvelle terminologie détaillée des autres observations et des spéculations sur le phénomène de l'érable à sucre sont proposées. [Traduit par la Rédaction]

Cité par
View all 2 citing articles
Lossy data structures

• Lossy data structures like Bloom Filters are a potential solution
• Bloom Filters allow you to test for set membership
• Instead of storing the object itself (String) they store a highly compressed bit signature
• One tailed error: never have false negatives, have false positives with some small, quantifiable probability
Harvesting data from the Web

- Mirror web sites
- Extract text page contents
- Perform language ID
- Segment into sentences
- Align document pairs
- Align sentences
- Remove duplicates
- ... Profit!
What I did

- 50M European Parliament
- 10^9 word webcrawl

1000M
French-English
10^9 word webcrawl
What Google does

Large Scale Parallel Document Mining for Machine Translation
Jakob Uszkoreit, Jay Ponte, Ashok Popat, Moshe Dubiner

2.5 billion general web pages
• Czech, English, French, German, Hungarian and Spanish

1.5 million OCRed public-domain books
• English, French and a few Spanish volumes
How is this different?

• How is the Google set-up different from mine?
• What resources and data do they have that I don’t?
• How do you think this might change their strategy?

• Discuss with your neighbor.
• Document translation pairs are simply near-duplicates, albeit annoyingly in different languages

• Use machine translation system to factor out differences in language and apply IR-inspired near duplicate detection techniques

• Pick-out small candidate sets of documents sharing a few rare matching features

• Score all pairs of documents in every candidate set using full features
Step 1: Translation

- Translate all input documents into a single language (e.g. English)
- Translation quality has only limited effect on data quality
- we’ll see that later in numbers
- Preprocess translations by removing stopwords and ‘boilerplate’ text
Step 2: Feature Extraction

• Extract 2 types of features from translated documents
• Matching features such that
  – Every translation pair is likely to have some of these features in common
  – Any given feature is unlikely to be shared by many documents
  – They use: 5-grams
• Scoring features
  – With higher overlap between the contents of two translations
  – Without frequency constraints
  – They use: bigrams
Step 2: Feature Extraction

• Generate two indexes
• Inverted index with every n-gram listing all document IDs with that n-gram
• Forward index with the set of scoring n-grams for each document
• (Embarrassingly parallel task)
Step 3: Prune Indexes

• Discard matching n-grams from inverted index
  – That are shared by more than a few (50) documents
  – That do not occur in more than one language

• Efficient operation on inverted index

• In parallel, annotate every occurrence of each scoring n-gram in the forward index with global information from the inverted index
  – Frequency
  – Number of original languages
  – Prune very frequent scoring n-grams (> 100,000 occurrences)
  – Prune scoring n-grams that occur only in one language
Step 4: Pairwise Scoring

• Get all pairs of document IDs that
  – share a given minimum number of matching n-grams
  – have similar lengths
  – are in two different, original languages

• Since frequent n-grams have been discarded, this generates relatively few candidate pairings and prevents $N^2$ explosion of comparisons

• Gather all candidate pairs for each document ID
Step 4: Pairwise Scoring

• Score candidate pairings and generating one n-best list per document, per language
  — Cosine similarity between idf n-gram vectors

• Further filter pairings by looking at relative order of shared n-grams

• (Again straightforward to parallelize -- Google loves that!)
Final Steps

• Discard pairings with scores below a threshold
• Discard pairings that are not symmetric
  – Document A is required to be in n-best list of document B and vice-versa
• Sentence-align the original documents using a standard dynamic programming algorithm
• Do lang ID and discard sentence pairs that are not detected to be in two different languages
• Discard those that with low IBM Model 1 probs
Number of words of mined English-foreign parallel text

<table>
<thead>
<tr>
<th>Language</th>
<th>baseline</th>
<th>books</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>27.5M</td>
<td>-</td>
<td>271.9M</td>
</tr>
<tr>
<td>French</td>
<td>479.8M</td>
<td>228.5M</td>
<td>4,914.3M</td>
</tr>
<tr>
<td>German</td>
<td>54.2M</td>
<td>-</td>
<td>3,787.6M</td>
</tr>
<tr>
<td>Hungarian</td>
<td>26.9M</td>
<td>-</td>
<td>198.9M</td>
</tr>
<tr>
<td>Spanish</td>
<td>441.0M</td>
<td>15.0M</td>
<td>4,846.8M</td>
</tr>
</tbody>
</table>

On the web data set, the system
• extracts 430 billion distinct 5-grams
• stores 500 billion bigram occurrences in forward index
• but performs less than 50 billion pairwise comparisons

Takes less than 24h on a cluster of 2,000 state-of-the-art CPUs
How much data did they get?

- Number of words of mined English-X parallel text

<table>
<thead>
<tr>
<th>Language</th>
<th>baseline</th>
<th>books</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>27.5M</td>
<td>-</td>
<td>271.9M</td>
</tr>
<tr>
<td>French</td>
<td>479.8M</td>
<td>228.5M</td>
<td>4,914.3M</td>
</tr>
<tr>
<td>German</td>
<td>54.2M</td>
<td>-</td>
<td>3,787.6M</td>
</tr>
<tr>
<td>Hungarian</td>
<td>26.9M</td>
<td>-</td>
<td>198.9M</td>
</tr>
<tr>
<td>Spanish</td>
<td>441.0M</td>
<td>15.0M</td>
<td>4,846.8M</td>
</tr>
</tbody>
</table>

- On the web data set, the system
  - extracts 430 billion distinct 5-grams
  - stores 500 billion bigram occurrences in forward index
  - but performs less than 50 billion pairwise comparisons

- Takes less than 24h on a cluster of 2,000 CPUs
How much did it improve their MT?

### Test Set 1

<table>
<thead>
<tr>
<th></th>
<th>baseline</th>
<th>+books</th>
<th>+web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech English</td>
<td>16.46</td>
<td>-</td>
<td>23.25 (+6.76)</td>
</tr>
<tr>
<td>German English</td>
<td>20.03</td>
<td>-</td>
<td>23.35 (+3.32)</td>
</tr>
<tr>
<td>Hungarian English</td>
<td>11.02</td>
<td>-</td>
<td>14.68 (+3.66)</td>
</tr>
<tr>
<td>French English</td>
<td>26.39</td>
<td>27.15 (+0.76)</td>
<td>28.34 (+1.95)</td>
</tr>
<tr>
<td>Spanish English</td>
<td>26.88</td>
<td>27.16 (+0.28)</td>
<td>28.50 (+1.62)</td>
</tr>
</tbody>
</table>

### Test Set 2

<table>
<thead>
<tr>
<th></th>
<th>baseline</th>
<th>+books</th>
<th>+web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech English</td>
<td>21.59</td>
<td>-</td>
<td>29.26 (+7.67)</td>
</tr>
<tr>
<td>German English</td>
<td>27.99</td>
<td>-</td>
<td>32.35 (+4.36)</td>
</tr>
<tr>
<td>French English</td>
<td>34.26</td>
<td>34.73 (+0.47)</td>
<td>36.65 (+2.39)</td>
</tr>
<tr>
<td>Spanish English</td>
<td>43.67</td>
<td>44.07 (+0.40)</td>
<td>46.21 (+2.54)</td>
</tr>
</tbody>
</table>
Google’s approach is great!

• Google’s approach is computational efficient and is embarrassingly simple to parallelize
• Generalizes across different types of documents
• Does not require presence of any metadata or document structure
• It employs many simple queries (matching n-grams)
• It has been applied to truly web-scale input data
• BUT there is a problem...
Problem: Everyone loves Google!

- There’s a problem: Google Translate is too good
- Everyone is using it to translate their web sites

- ... So Google ends up harvesting its own translations as parallel corpora to train its system!
- When they train a new version of the system it reverts back to behaving like the old version
Solution: Digital Watermarking
Watermarking SMT output

Watermarking the output of Structured Prediction with an application in Statistical Machine Translation

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz J. Och, Juri Ganitkevitch

“Back-of-the-envelope” study:

Corpora identified by Uszkoreit et al 2010

\[ \cap \]

Pages using translate plugins to serve content in multiple languages

<table>
<thead>
<tr>
<th>Language pair</th>
<th>% in set / all identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagalog-English</td>
<td>50.6%</td>
</tr>
<tr>
<td>Hindi-English</td>
<td>44.5%</td>
</tr>
<tr>
<td>Galician-English</td>
<td>41.9%</td>
</tr>
</tbody>
</table>
Task: Identify One’s Own MT output

**Assumption**: each translation output has $k$ relatively similar alternatives

![Diagram showing the selection process](image)

**Intuition**: rather than simply selecting the “best” translation according to the model, systematically select alternative results such that we can identify them.
Watermarking Selection

\[ r' = \underset{r \in D_k(q)}{\text{argmax}} \; w(r, D_k(q), h) \]

• \( r \): the machine translated output sentence
• \( h \): a random hash function
• \( w \): a selector function to choose from the set of \( k \) alternatives
Watermarking Evaluation

• **False Positive Rate**: how often are non-watermarked collections falsely identified as watermarked

• **Recall Rate**: how often watermarked collections are correctly identified as watermarked

• **Quality Degradation**: how does the selected translation differ from best translation under BLEU?
Random Hashing

A good $h$ produces independent bits, implying the number of #1s:

$$\chi \sim \text{Binomial}(p = 0.5, n = |h(C_n)|)$$
Random Hashing

Null Hypothesis: an un-marked collection would generate bit sequences where #1s follows:

\[ X \sim Binomial(p = 0.5, n = |h(C_n)|) \]
Systematically Selecting Improbable Results

$q$

\[ D_k(q) \]

0011...1001
1111...1101
0011...1001

Improbable result lots more 1s.
### Evaluation: False Positive Rates

<table>
<thead>
<tr>
<th>Language</th>
<th>False Positive Rate: full sentences: %</th>
<th>False Positive Rate: using 3-5 grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>2.4</td>
<td>5.8</td>
</tr>
<tr>
<td>French</td>
<td>1.8</td>
<td>7.5</td>
</tr>
<tr>
<td>Hindi</td>
<td>5.6</td>
<td>3.5</td>
</tr>
<tr>
<td>Turkish</td>
<td>5.5</td>
<td>6.2</td>
</tr>
</tbody>
</table>

BLEU loss can be held to -0.2 for most languages
Evaluation: Bound at -0.2 BLEU Loss

The chart shows the recall for various languages: Arabic, French, Hindi, and Turkish. The recall values are measured at different levels: sentence-level and 3-to-5 grams.
• On several languages it is possible to achieve:
  – high recall rates (over 80%)
  – low false positive rates (5-8%)
  – minimal quality degradation (-0.2 BLEU)
  – allowing for local edit operations

• Problem solved! Your TA is a hero!
Questions?