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!

Sentence aligned bitexts

Arabic	English
فالتعذيب لا يزال يمارس على نطاق واسع	Torture is still being practised on a wide scale.
وتتم عمليات الاعتقال والاحتجاز دون سبب بصورة روتينية	Arrest and detention without cause take place routinely.
وحان وقت التحلى بالبصيرة والشجاعة السياسية .	This is a time for vision and political courage
. . .	. . .

Chinese	English
我国 能源 原材料 工业 生产 大幅度 增长 .	China's energy and raw materials production up.
非国大 要求 阻止 更 多 被 拘留 人员 死亡 .	ANC calls for steps to prevent deaths in police custody .
. . .	. . .

Goals for today's lecture

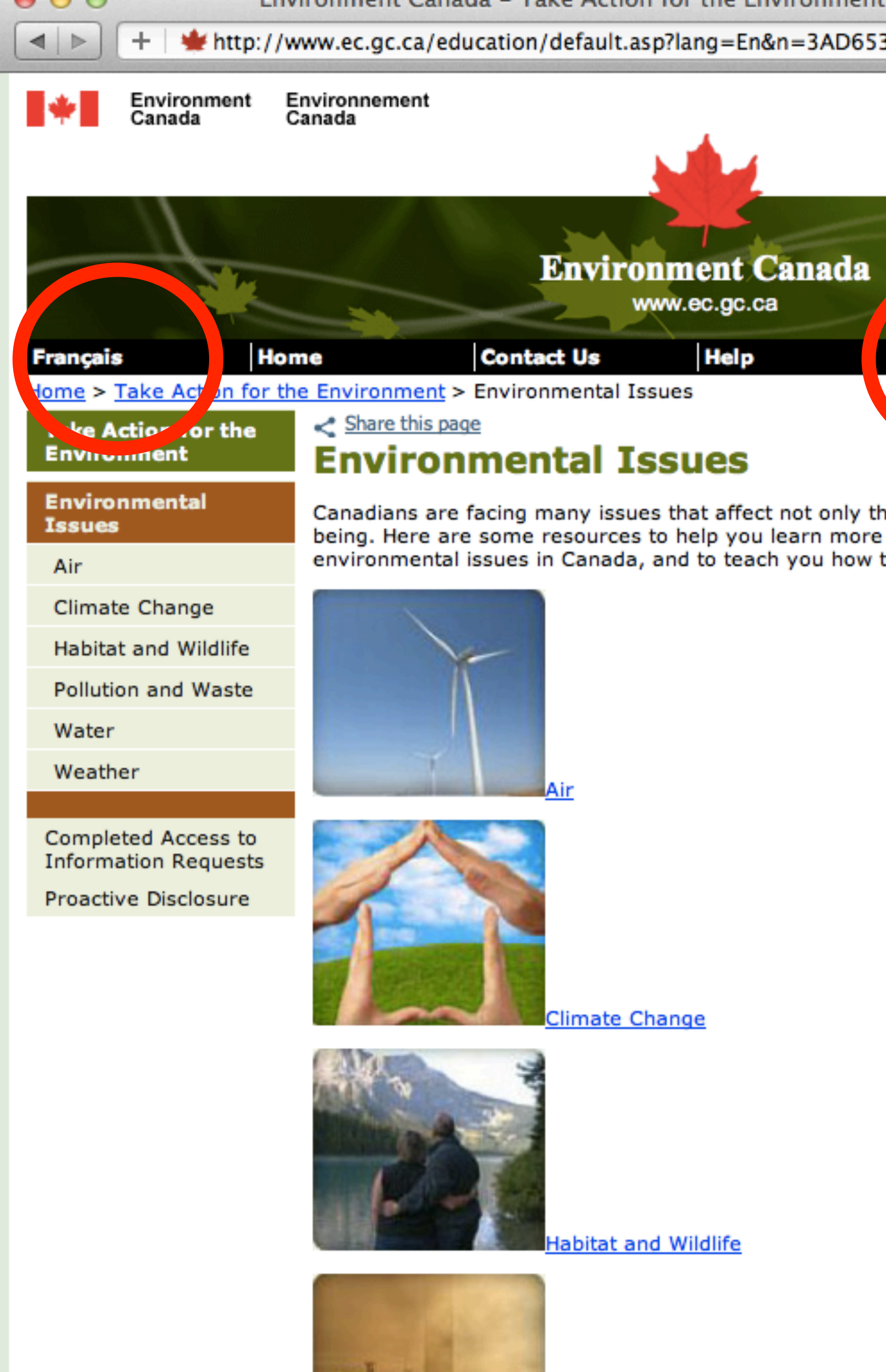
- Understand how to mine bitexts from the web
- Web Crawling 101
- Review recent research into extracting parallel documents from the web and from unstructured collections
- What to do if you're Google and you're worried about harvesting your own machine translation output

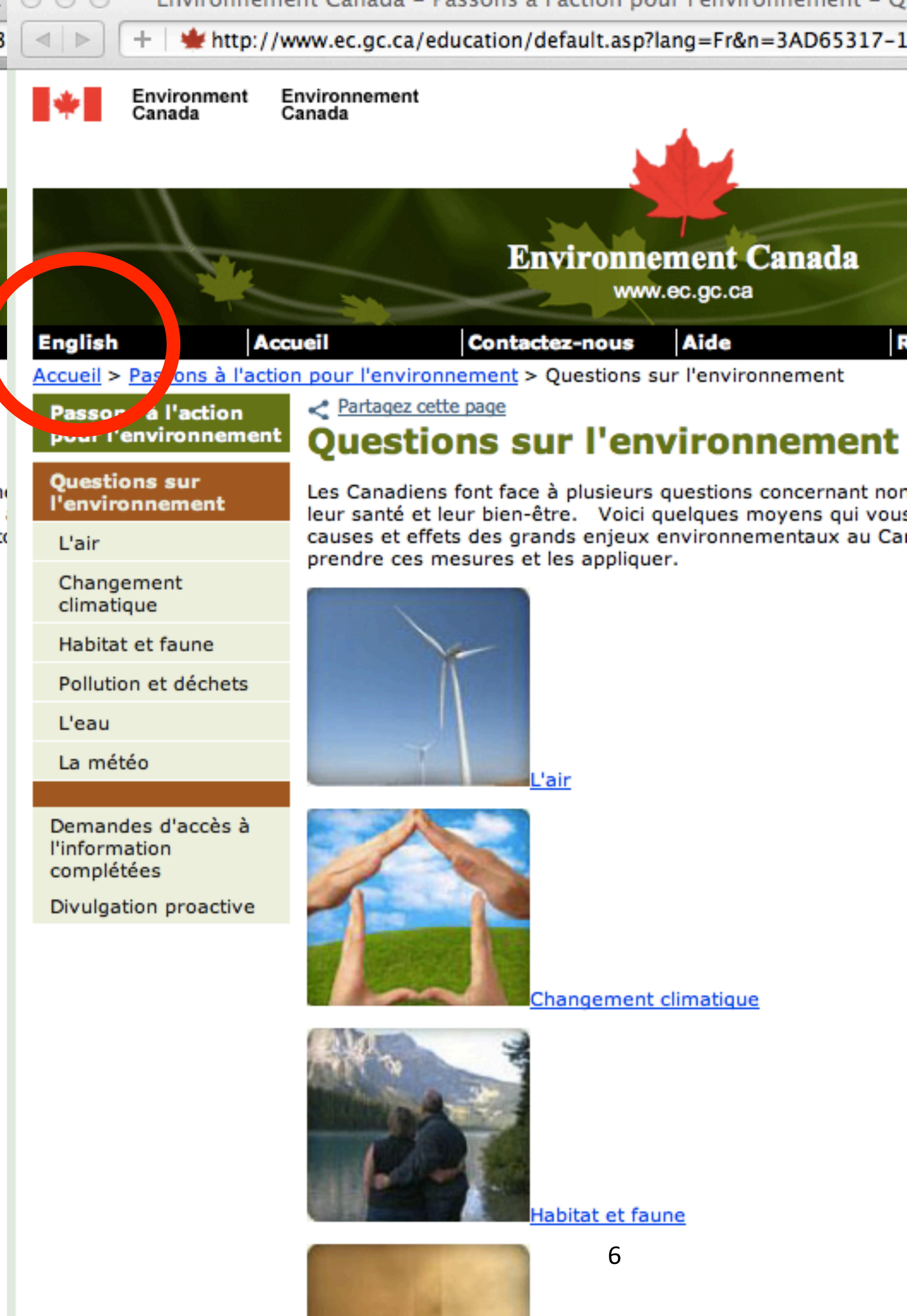
The Web as a Parallel Corpus

- Old idea:
  - Philip Resnik, "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text", in Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98), October, **1998.**
- 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 Français
- 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


http://www.ec.gc.ca/education/default.asp?lang=En&n=3AD653
Environment Canada
Environnement Canada
Environment Canada
www.ec.gc.ca
Français
Home
Contact Us
Help
Home > Take Action for the Environment > Environmental Issues
Take Action for the Environment
Share this page
Environmental Issues
Environmental Issues
Air
Climate Change
Habitat and Wildlife
Pollution and Waste
Water
Weather
Completed Access to Information Requests
Proactive Disclosure
Canadians are facing many issues that affect not only th
being. Here are some resources to help you learn more
environmental issues in Canada, and to teach you how to
Air
Climate Change
Habitat and Wildlife


http://www.ec.gc.ca/education/default.asp?lang=Fr&n=3AD65317-1
Environment Canada
Environnement Canada
Environnement Canada
www.ec.gc.ca
English
Accueil
Contactez-nous
Aide
Accueil > Passons à l'action pour l'environnement > Questions sur l'environnement
Passons à l'action pour l'environnement
Partagez cette page
Questions sur l'environnement
Questions sur l'environnement
L'air
Changement climatique
Habitat et faune
Pollution et déchets
L'eau
La météo
Demandes d'accès à l'information complétées
Divulgation proactive
Les Canadiens font face à plusieurs questions concernant nor
leur santé et leur bien-être. Voici quelques moyens qui vous
causes et effets des grands enjeux environnementaux au Ca
prendre ces mesures et les appliquer.
L'air
Changement climatique
Habitat et faune

Wikipedia
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

Interaction

Toolbox

Print/export
- Create a book
- Download as PDF
- Printable version

Languages
- Euskara
- Français

Article | Talk

Pyrenean G

From Wikipedia, the fre

Not to be confused

The **Pyrenean** goat b
Pyrenees of France a
Cantabrian Mountains
the production of milk

Sources

- Pyrenean Goat

This goat-re
stub. You ca
expanding it

Rate this pag
What's this?

Trustworthy

I am highly kno

Categories: Goat br
Goat breeds origina

This page was last modi

Text is available under t
of use for details.
Wikipédia® is a registere

Wikipédia
L'encyclopédie libre

- Accueil
- Portails thématiques
- Index alphabétique
- Article au hasard
- Contacter Wikipédia

Contribuer
- Premiers pas
- Aide
- Communauté
- Modifications récentes
- Faire un don

Imprimer / exporter

Boîte à outils

Autres langues
- English
- Euskara

Article | Discussion | Lire | Modifier | Afficher l'historique | Rechercher

Pyrénées (race caprine)

Pour les articles homonymes, voir Pyrénées (homonymie).

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 précoce du chevreau à 2 mois 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 médiocre, dans des conditions climatiques parfois très rudes. Elle permet de maintenir certains paysages ouverts en empêchant qu'ils ne s'embroussaillent.

Pyrénées

Chèvre pyrénéenne

Espèce	Chèvre (*Capra aegagrus hircus*)
Région d'origine	
Région	Pyrénées, France
Caractéristiques	
Taille	Grande
Robe	Brune ou noire avec des taches blanches
Autre	
Diffusion	Locale
Utilisation	Lait et viande

modifier

Sommaire [masquer]

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 ⋮	<HTML> <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]	
[Chunk:13]	
[END:H1]	
[Chunk:112]	[Chunk:122]

STRAND

- % 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)

URL similarity

www.aecb.org/**fra**/publisher.asp?id=4090
www.aecb.org/**eng**/publisher.asp?id=4090

porta
porta

www
www

www
www

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/

www.rwanda-botschaft.de/embassy3/pages/341763a3c5e7f86ced395a8f0e32b8d7nw.php?**lg=fr**&src=ns0000501151840&nId=44&diflg=nodif
www.rwanda-botschaft.de/embassy3/pages/

Sites with translated content

93236 rparticle.web-p.cisti.nrc.ca
53973 www.ec.gc.ca
52318 www.hc-sc.gc.ca
45118 portal.unesco.org
42737 www.cra-arc.gc.ca
34617 www.dfo-mpo.gc.ca
29445 www.canadianheritage.gc.ca
28170 www.idrc.ca
26823 www.agr.gc.ca
21255 www.dfait-maeci.gc.ca
19827 www.forces.gc.ca
16922 www.ic.gc.ca
16492 www.ceaa-acee.gc.ca
16289 www.gg.ca
15002 www.canadianencyclopedia.ca

14380 www2.parl.gc.ca
14089 www.fin.gc.ca
13706 www.aecb.org
13264 www.cihr-irsc.gc.ca
12161 www.cprn.org
12145 www.civilisations.ca
11632 www.cbsa.gc.ca
11632 www.cbsa-asfc.gc.ca
11005 www.hockeycanada.
10382 www.crr.ca
10338 www.commonlaw.uot
10150 www.ourroots.ca
9224 www.cws-scf.ec.gc.ca
8440 www.elections.ca
8099 www.collectionscanad

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/
```

No robots

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:

  wget **-robots=off** --mirror http://akhbarlive.com/

- Some sites will block wget directly, you can pretend to be some other browser:

  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:

  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

Extract text content

- 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

Perform language ID

- 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!

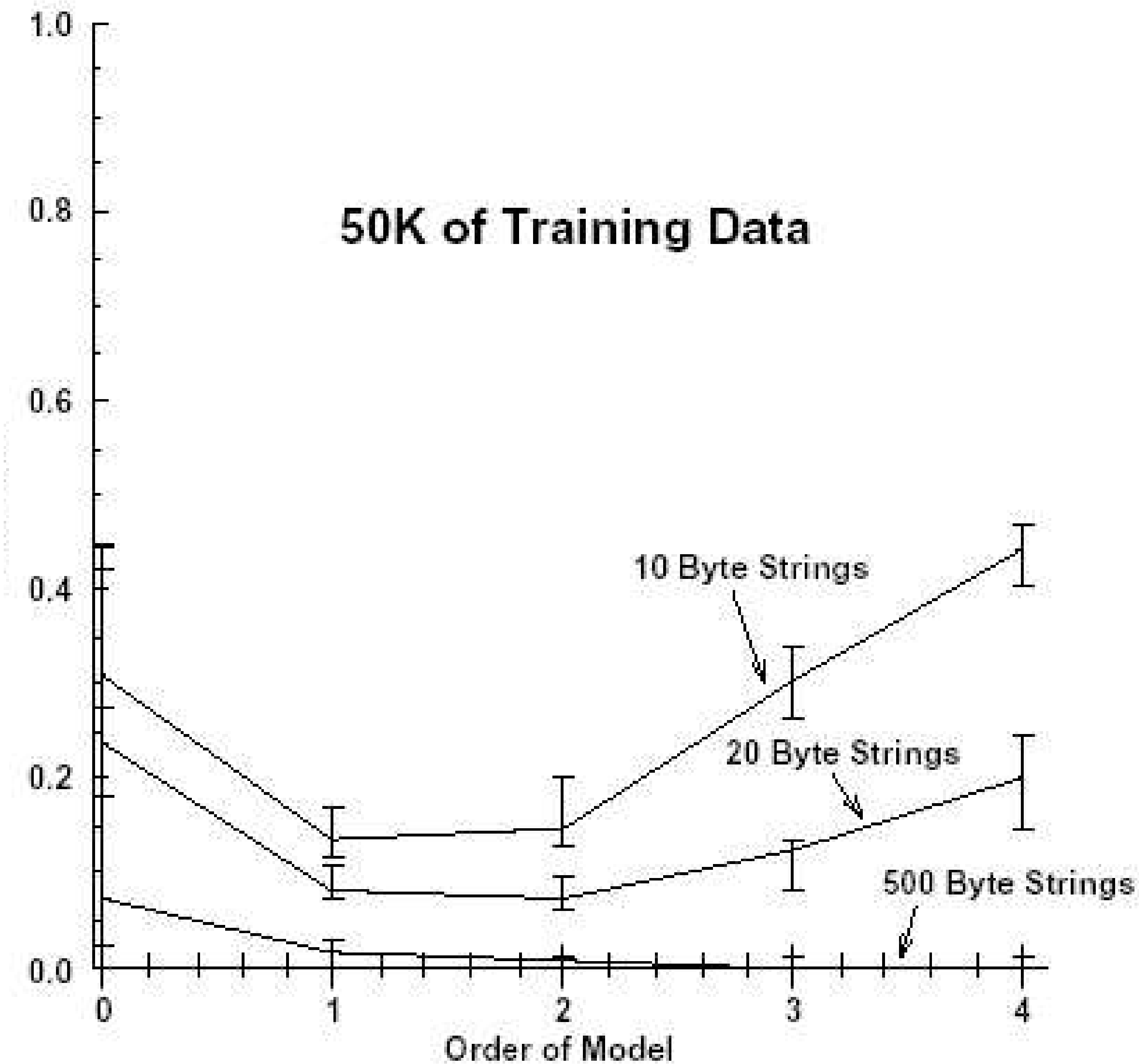
Statistical language ID

- Intuition: some character strings are more probable in one language than in others

Language	char sequence
Dutch	*vnd*
English	*ery*
French	*eux*
Gaelic	*mh*
German	*der*
Italian	*cchi*
Portuguese	*seu*
Serbo-croat	*lj*
Spanish	*ir*

Dunning (1994)

$$p(S \mid A) = p(s_1 \ldots s_k \mid A) \prod_{i=k+1}^{N} p(s_i \mid s_{i-k} \ldots s_{i-1} \mid A)$$



Segment into sentences

- 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

Align document pairs

- 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/aafd5dcf-4dcc-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

Subscriber access provided by JOHNS HOPKINS UNIVERSITY

Login Regis

110 YEARS OF SCIENCE IN JUST ONE CLICK

Journal of Forest Research

oks Authors Librarians Societies About Us Contact Help

of Forest Research > List of Issues > Volume 29, Number 11, December 1999 > The birdseye figured grain in sugar maple

Article

TOC Next »

The birdseye figured grain in sugar maple (*Acer saccharum*): literature review, nomenclature, and structural characteristics

Don C. Bragg

PDF (9052 K)

PDF-Plus (1551 K)

Also read

Citing articles

Canadian Journal of Forest Research, 1999, 29(11): 1637-1648, 10.1139/x99-155

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, fingernail, 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

Accès de l'abonné offert par JOHNS HOPKINS UNIVERSITY

110 YEARS OF SCIENCE IN JUST ONE CLICK

nadienne de recherche foresti

Livres Auteurs Bibliothécaires Sociétés À propos de nous

nne de recherche forestière > Liste de numéros > Volume 29, numéro 11, décembre 199

Article

Tal

The birdseye figured grain in sugar maple (*Acer saccharum*): literature review, nomenclature, and structural characteristics

Don C. Bragg

Revue canadienne de recherche forestière, 1999, 29(11): 1637-1648, 10.1

RÉSUMÉ

On connaît peu de chose à propos du grain de l'érable à sucre (*Acer sac*
présence des mouchetures. Cet article clarifie et élargit la discussion au
décrivant les similitudes et les différences avec le grain texturé chez d'au
qu'en discutant des caractéristiques importantes de son anatomie partic
des sections consacrées à la discussion des causes possibles de l'érabl
géographique détaillée de l'érable piqué et à ce qu'on connaît du rôle de
l'érable piqué. Les variations possibles de la moucheture typique (p. ex.,
chat, déformée) sont présentées et une nouvelle terminologie descriptive
d'autres observations et spéculations sur le phénomène de l'érable piqu
orientations de recherche sont proposées.[Traduit par la Rédaction]

Cité par

View all 2 citing articles

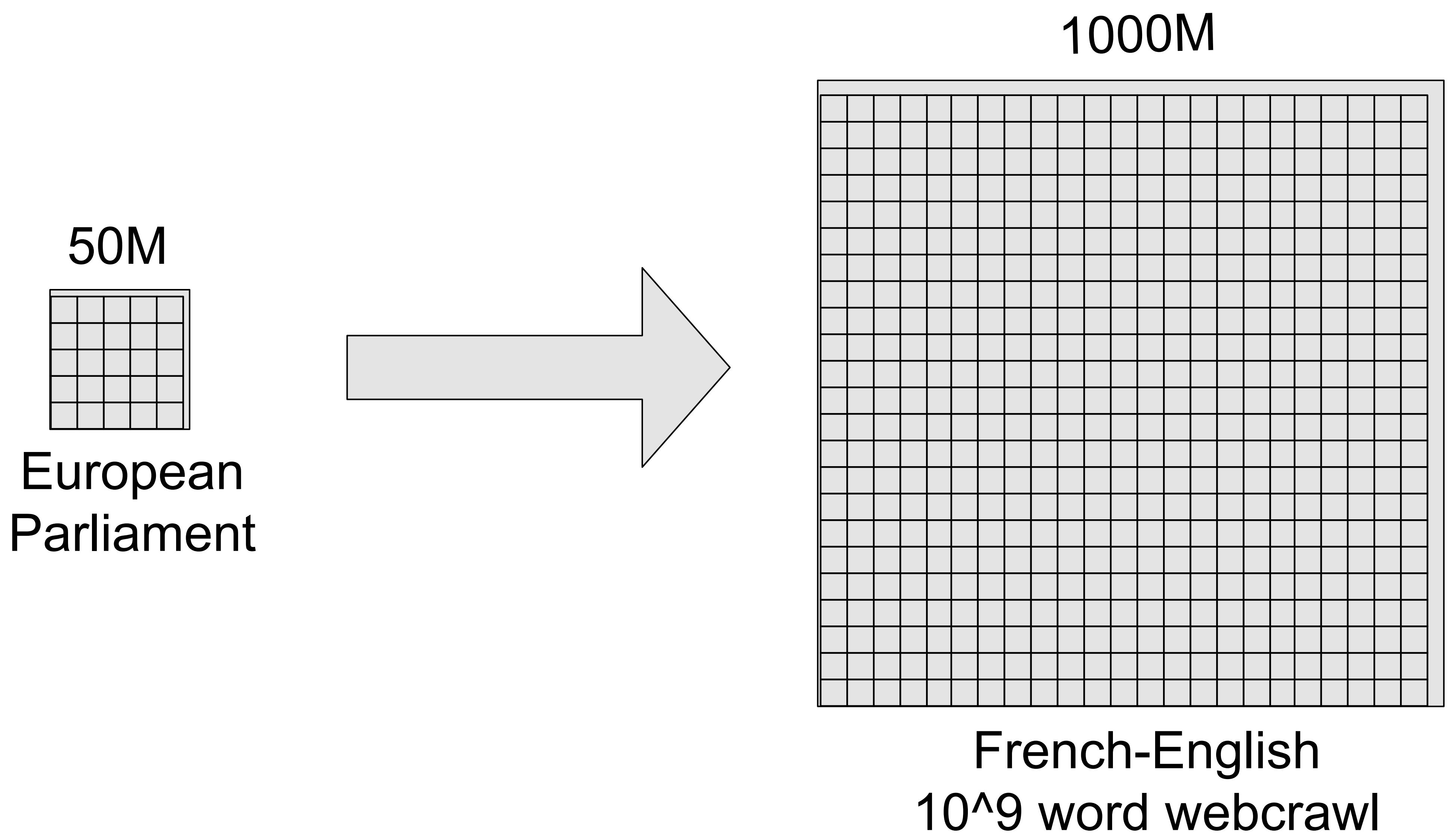
Lossy data structures

- 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
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.

High level strategy

- 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

	baseline	books	web
Czech	27.5M	-	271.9M
French	479.8M	228.5M	4,914.3M
German	54.2M	-	3,787.6M
Hungarian	26.9M	-	198.9M
Spanish	441.0M	15.0M	4,846.8M

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

	baseline	books	web
Czech	27.5M	-	271.9M
French	479.8M	228.5M	4,914.3M
German	54.2M	-	3,787.6M
Hungarian	26.9M	-	198.9M
Spanish	441.0M	15.0M	4,846.8M

- 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

	baseline	+books	+web
Czech English	16.46	-	23.25 (+6.76)
German English	20.03	-	23.35 (+3.32)
Hungarian English	11.02	-	14.68 (+3.66)
French English	26.39	27.15 (+0.76)	28.34 (+1.95)
Spanish English	26.88	27.16 (+0.28)	28.50 (+1.62)

Test Set 2

	baseline	+books	+web
Czech English	21.59	-	29.26 (+7.67)
German English	27.99	-	32.35 (+4.36)
French English	34.26	34.73 (+0.47)	36.65 (+2.39)
Spanish English	43.67	44.07 (+0.40)	46.21 (+2.54)

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 Uskzoreit et al 2010

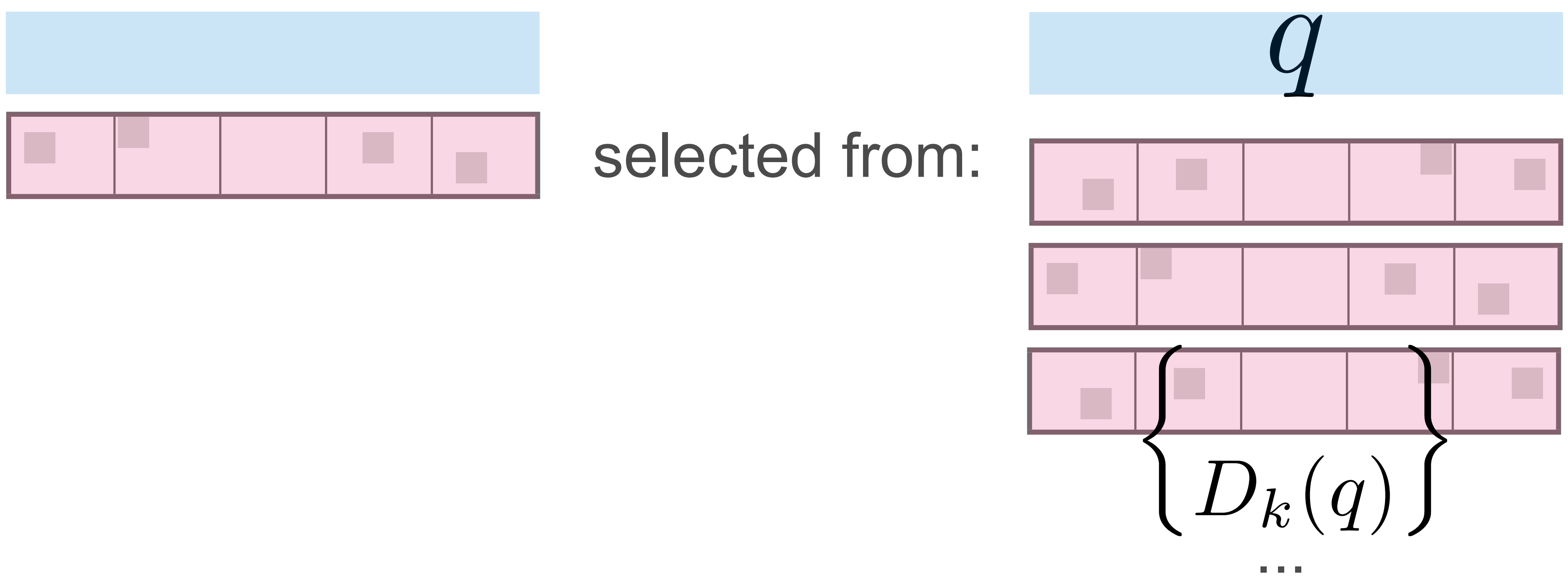
$\cap$

Pages using translate plugins to serve content in multiple languages

Language pair	% in set / all identified
Tagalog-English	50.6%
Hindi-English	44.5%
Galician-English	41.9%

Task: Identify One's Own MT output

Assumption: each translation output has k relatively similar alternatives



Intuition: rather than simply selecting the "best" tranlsation according to the model, systematically select alternative results such that we can identify them.

Watermarking Selection

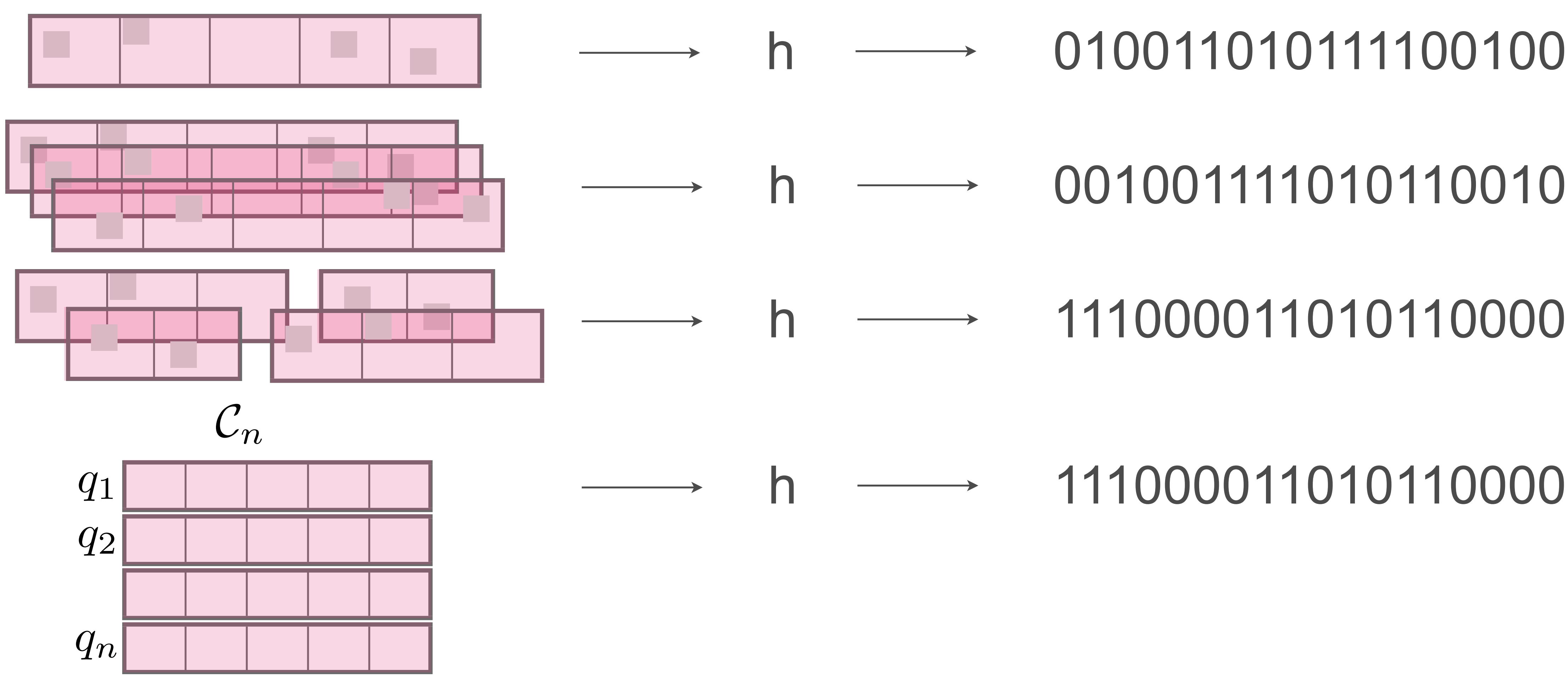
$$r' = \underset{r \in D_k(q)}{argmax} \; w(r, D_k(q), h)$$

- 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?

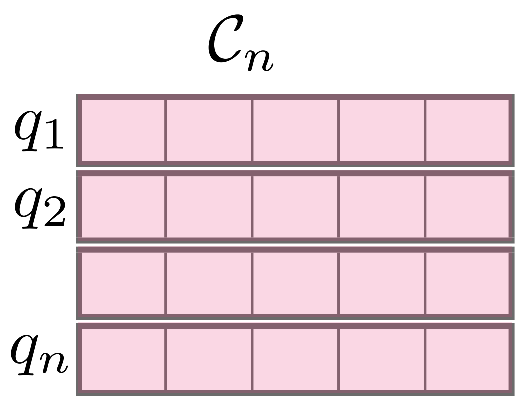
Random Hashing



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

$$\mathcal{X} \sim Binomial(p = 0.5, n = |h(\mathcal{C}_n)|)$$

Random Hashing

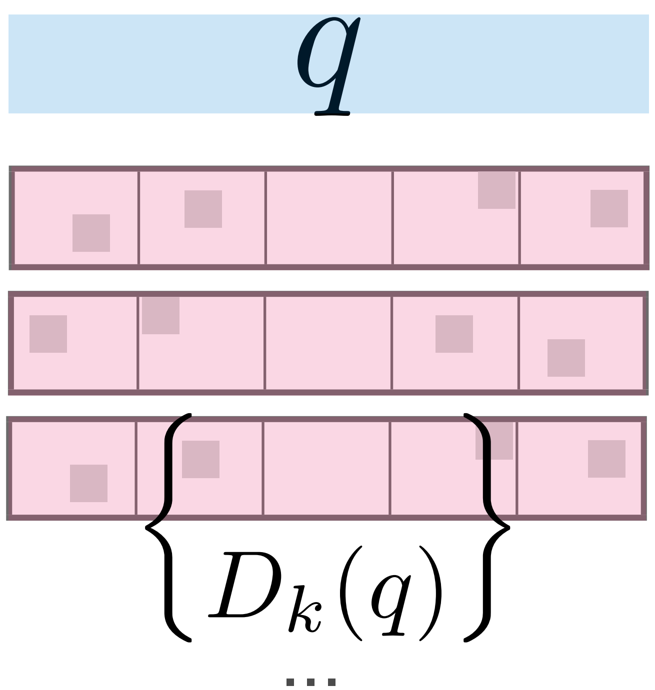


$\longrightarrow$ h $\longrightarrow$ 111000011010110000

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

$$\mathcal{X} \sim Binomial(p = 0.5, n = |h(\mathcal{C}_n)|)$$

Systematically Selecting Improbable Results



0011...1001

1111...1101 ← Improbable result lots more 1s.

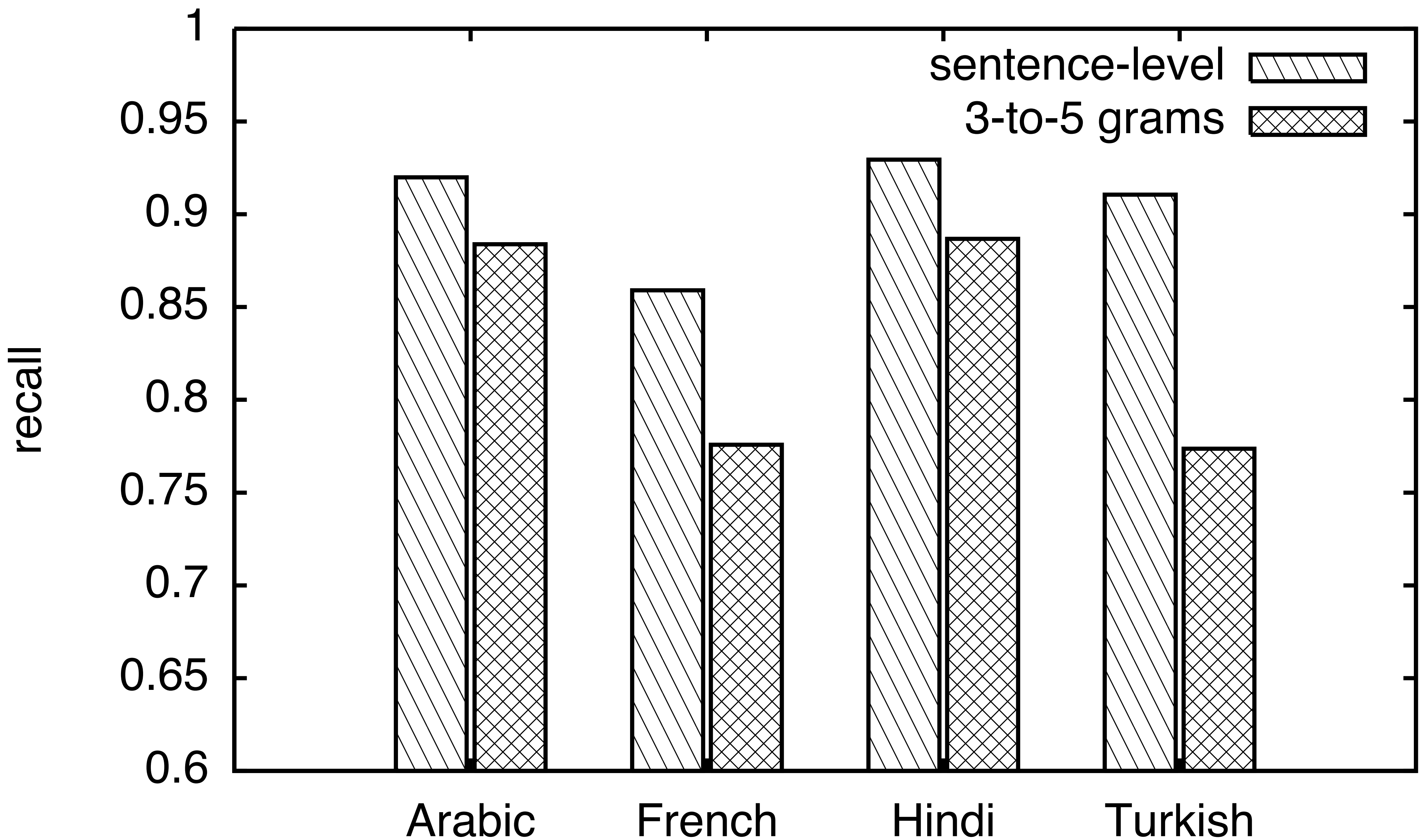
0011...1001

Evaluation: False Positive Rates

Language	False Positive Rate: full sentences: %	False Positive Rate: using 3-5 grams
Arabic	2.4	5.8
French	1.8	7.5
Hindi	5.6	3.5
Turkish	5.5	6.2

BLEU loss can be held to -0.2 for most languages

Evaluation: Bound at -0.2 BLEU Loss



Watermarking wrap up

- 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?