Beyond Translation Memories

Finding similar documents in comparable corpora

Project TTC
Terminology Extraction, Translation Tools and Comparable Corpora
(FP-7, STREP-248.005)

30 November 2012

Outline

1. Parallel and comparable resources
   - Cline of comparability
   - Collecting comparable corpora

2. Measuring text similarity
   - Features to compare texts
   - Anchor-based method

Cline of comparability

Parallel texts Traditional source of texts in MT:

Cline of comparability

Parallel texts Traditional source of texts in MT:

- true and accurate translations;

Cline of comparability

Parallel texts Traditional source of texts in MT:

- true and accurate translations;
- reasonable translations with minor language-specific variations (*New York* → 北京);

Cline of comparability

Parallel texts Traditional source of texts in MT:

- true and accurate translations;
- reasonable translations with minor language-specific variations (*New York* → 北京);

Strongly comparable texts Heavily edited translations or independent, but closely related texts:

Cline of comparability

Parallel texts Traditional source of texts in MT:

- true and accurate translations;
- reasonable translations with minor language-specific variations (*New York* → 北京);

Strongly comparable texts Heavily edited translations or independent, but closely related texts:

- same source with same editorial control (BBC News in English and Romanian);

Cline of comparability

Parallel texts Traditional source of texts in MT:

- true and accurate translations;
- reasonable translations with minor language-specific variations (*New York* → 北京);

Strongly comparable texts Heavily edited translations or independent, but closely related texts:

- same source with same editorial control (BBC News in English and Romanian);
- related independently written texts on same topic (interlinked Wikipedia articles, same event from AFP, DPA and Reuters);

Cline of comparability

Weakly comparable texts More variation in topics

English and Chinese textbooks on designing the wind turbines, or parliamentary debates on health care from the Bundestag, the House of Commons and the Russian Duma;

Cline of comparability

Weakly comparable texts More variation in topics

English and Chinese textbooks on designing the wind turbines, or parliamentary debates on health care from the Bundestag, the House of Commons and the Russian Duma;

General unrelated texts Texts representing language varieties without a claim to their relatedness

Brown corpus, LOB, Lancaster Corpus of Mandarin Chinese (A. Press: reportage, B. Press: editorial, C. Press: Reviews. . .); Web snapshots in Chinese, German and Italian

Existing studies

- Parallel fragments for MT
  parallel sentences in Wikipedia articles [Adafre and de Rijke, 2006]
  BBC News in English and Romanian [Munteanu and Marcu, 2006]
  CLIR in news corpora [Zhao and Vogel, 2002]

Existing studies

- Parallel fragments for MT
  parallel sentences in Wikipedia articles [Adafre and de Rijke, 2006]
  BBC News in English and Romanian [Munteanu and Marcu, 2006]
  CLIR in news corpora [Zhao and Vogel, 2002]
- Terminology extraction
  Dictionary-based methods
  [Li and Gaussier, 2010, Su and Babych, 2012]

Rationale

- Limited parallel resources for new domains (mostly institutional repositories, like UN or EuroParl)

Rationale

- Limited parallel resources for new domains (mostly institutional repositories, like UN or EuroParl)
- Noisy comparable corpora: different topics

Rationale

- Limited parallel resources for new domains (mostly institutional repositories, like UN or EuroParl)
- Noisy comparable corpora: different topics
- Few or no dictionaries

Rationale

- Limited parallel resources for new domains (mostly institutional repositories, like UN or EuroParl)
- Noisy comparable corpora: different topics
- Few or no dictionaries
- Purpose: Terminology Extraction

Collecting weakly comparable corpora

Three strategies:

1. targeted crawling of specific resources, which are known to be comparable;

Collecting weakly comparable corpora

Three strategies:

1. targeted crawling of specific resources, which are known to be comparable;
2. collection of responses from search engines using parallel terms;

Collecting weakly comparable corpora

Three strategies:

1. targeted crawling of specific resources, which are known to be comparable;
2. collection of responses from search engines using parallel terms;
3. focused crawling which starts from a small number of seeds.

Targeted crawling

Language	id	Million Tokens	Articles	EN iwikis
Chinese	zh	101	137179	87389
Dutch	nl	163	435716	290979
English	en	1440	2524134	n/a
French	fr	459	838771	541715
German	de	563	1114696	603437
Portuguese	pt	156	361204	245102
Russian	ru	268	609525	345195
Spanish	es	365	664097	438864
Ukrainian	uk	81	214403	139827

Crawling Wikipedia

```
’’’Biogas’’’ typically refers to a [[gas]] produced by
breakdown of [[organic matter]] in the absence of [[oxygen]].
...
[[Category:Anaerobic digestion]]
[[Category:Biofuels]]
[[Category:Biodegradation]]
...
[[af:Biogas]]
[[bg:Биогаз]]
[[ca:Biogàs]]
[[cs:Bioplyn]]
[[da:Biogas]]
[[de:Biogas]]
[[et:Biogaas]]
```

English-Russian energy corpus from Wikipedia

	Wiki-En		Wiki-Ru	
	texts	words	texts	words
Nuclear	158	254014	165	130797
Renewable	51	99960	51	47927

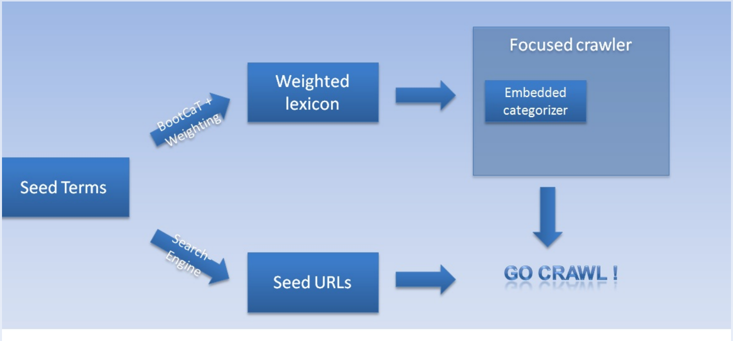
Keywords for comparable corpus

wind farm	风力发电厂	ветроэлектростанция
geothermal power	地热能	геотермальная энергия
hydroelectricity	水力发电	гидроэнергетика
photovoltaics	太阳能光伏	фотоэлектричество

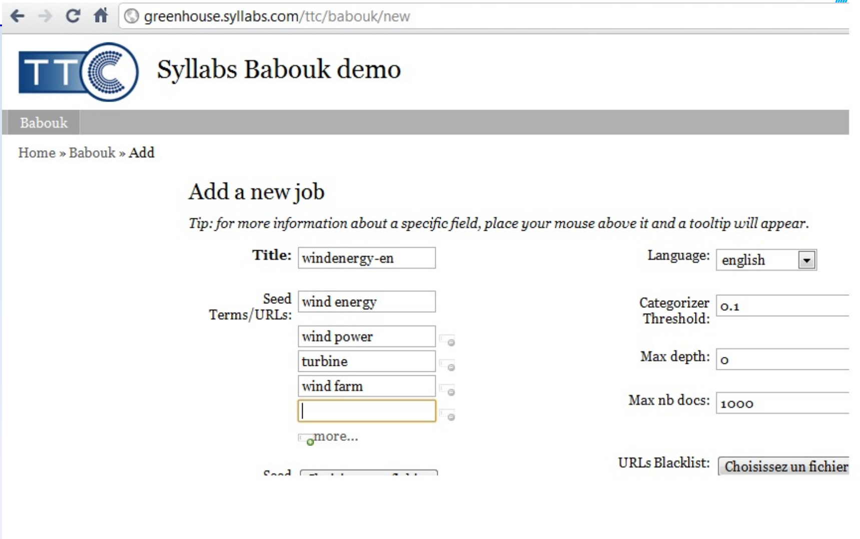
...

	Crawled-En		Crawled-Ru		Crawled-Zh	
	texts	words	texts	words	texts	words
Renewable	5762	7505765	5126	7766462	3287	12431752

Focused crawling



Focused crawling interface



greenhouse.syllabs.com/ttc/babouk/new

Syllabs Babouk demo

Home » Babouk » Add

Add a new job

Tip: for more information about a specific field, place your mouse above it and a tooltip will appear.

Title: windenergy-en

Seed Terms/URLs: wind energy

wind power

turbine

wind farm

Language: english

Categorizer Threshold: 0.1

Max depth: 0

Max nb docs: 1000

URLs Blacklist: Choisissez un fichier

Outline

1 Parallel and comparable resources
- Cline of comparability
- Collecting comparable corpora

2 Measuring text similarity
- Features to compare texts
- Anchor-based method

Features to compare texts

- 500 most frequent words [Kilgarriff, 2001]

Features to compare texts

- 500 most frequent words [Kilgarriff, 2001]
- *hapax legomena* [Patry and Langlais, 2011] or named entities [Goeuriot et al., 2009]

Features to compare texts

- 500 most frequent words [Kilgarriff, 2001]
- *hapax legomena* [Patry and Langlais, 2011] or named entities [Goeuriot et al., 2009]
- flexigrams [Forsyth and Sharoff, 2011]
  oil and gas industry → *oil industry*, *gas industry*, *oil and*, *and gas*, *oil gas*, *and industry*

Features to compare texts

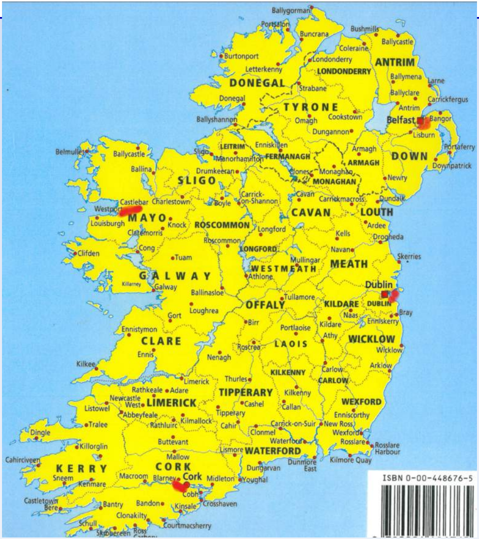
- 500 most frequent words [Kilgarriff, 2001]
- *hapax legomena* [Patry and Langlais, 2011] or named entities [Goeuriot et al., 2009]
- flexigrams [Forsyth and Sharoff, 2011]
  oil and gas industry → *oil industry*, *gas industry*, *oil and*, *and gas*, *oil gas*, *and industry*
- part-of-speech signatures [Sharoff, 2010, Santini et al., 2006] or mix of frequent words with POS tags
  [Baroni and Bernardini, 2006, Sharoff, 2007] for genres

Keywords for ‘Darrieus wind turbine’

LL	DocF	Word	Total
326.11	15	Darrieus	50
224.29	21	blade	20269
208.15	19	turbine	15976
135.47	18	wind	84724
95.85	9	torque	8821
79.80	6	aerofoil	1559
68.83	9	angle	39843
66.15	17	design	516317
61.91	6	rotor	6944
54.37	7	spin	29185
50.57	8	generate	69293
47.59	6	rotate	23012
46.95	9	speed	137005
35.88	4	propeller	9107
35.74	2	self-starting	52
34.51	6	pitch	69433
34.08	3	airflow	2059
33.51	10	force	405295
32.11	6	tower	85230
25.84	4	load	32271
25.74	4	conventional	32674
25.48	4	vertical	33781

LL	DocF	Word	Total
145.18	10	ротор (rotor)	1889
130.67	9	дарье (Darrieus)	1699
65.57	3	самозапуск (self-start)	13
56.41	6	поток (flow)	14719
51.51	6	крыло (aerofoil)	22180
34.36	2	подъёмная (lifting)	99
33.39	3	турбина (turbine)	3092
29.94	3	вектор (vector)	5503
27.29	4	скорость (speed)	35960
24.16	2	мгновенный (instant)	1280
18.47	4	сила (force)	111062
17.67	2	вращение (spin)	6518
17.25	2	плохой (difficult)	7239
17.25	2	коэффициент (rate)	7253

Geometric interpretation of anchoring



Separating Ireland

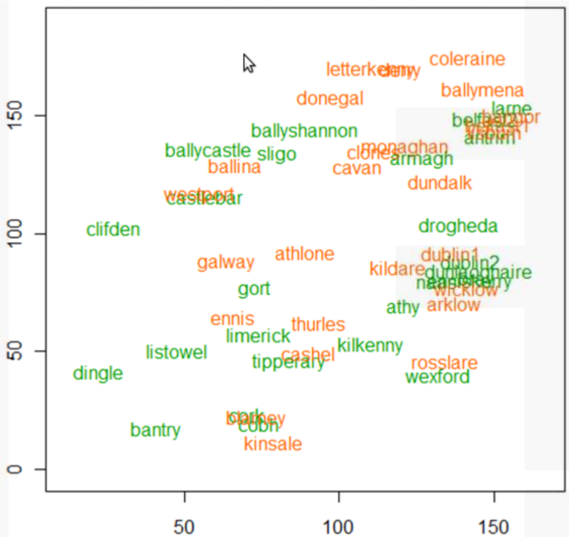
Town	antrim	armagh	athy
antrim	0	31	105
armagh	31	0	74
athy	105	74	0

. . .

Town	arklow	athlone	ballina
arklow	0	80	146
athlone	80	0	66
ballina	146	66	0

26 dimensions in each list

Re-uniting Ireland



(Spearman's) Rank correlation: 0.9941–0.9986

Procedure for texts

1. Select/inject anchors

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)
   - generic parallel texts (5g corpus)

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)
   - generic parallel texts (5g corpus)

NB $N_a << N_c$

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)
   - generic parallel texts (5g corpus)

   NB $N_a << N_c$
2. Compute monolingual distances to anchors

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)
   - generic parallel texts (5g corpus)

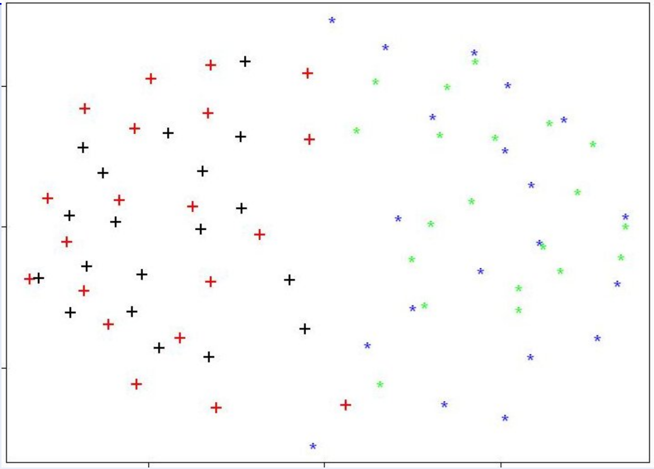
   NB $N_a << N_c$
2. Compute monolingual distances to anchors
3. Compute distances between texts represented by their distances to anchors

Procedure for texts

1. Select/inject anchors
   - in-domain parallel texts
   - in-domain comparable texts (manual or automatic)
   - generic parallel texts (5g corpus)

   NB $N_a << N_c$
2. Compute monolingual distances to anchors
3. Compute distances between texts represented by their distances to anchors
4. Perform MDS (Multi-Dimensional Scaling) for visualisation purposes

Comparability across languages



+=RuNu, +=EnNu, *=RuRe, *=EnRe

Term alignment

- Similarity between keywords of related documents [Rapp et al., 2012]

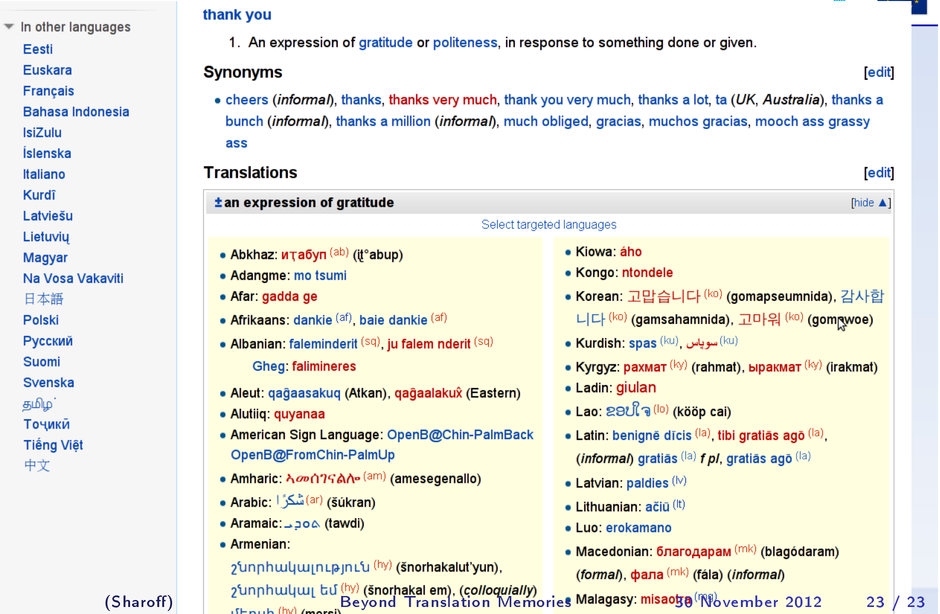
Term alignment

- Similarity between keywords of related documents [Rapp et al., 2012]
- Initial weights for linking terms:
  airflow → поток, Дарье, самозапуск, лопасть
  blade → поток, Дарье, самозапуск, лопасть
  Darrieus → поток, Дарье, самозапуск, лопасть
  self-starting → поток, Дарье, самозапуск, лопасть

Term alignment

- Similarity between keywords of related documents [Rapp et al., 2012]
- Initial weights for linking terms:
  airflow → поток, Дарье, самозапуск, лопасть
  blade → поток, Дарье, самозапуск, лопасть
  Darrieus → поток, Дарье, самозапуск, лопасть
  self-starting → поток, Дарье, самозапуск, лопасть
- Spreading activation
  airflow → поток
  Darrieus → Дарье
  blade → лопасть
  self-starting → самозапуск

Thank you from Wiktionary



In other languages
- Eesti
- Euskara
- Français
- Bahasa Indonesia
- IsiZulu
- Íslenska
- Italiano
- Kurdî
- Latviešu
- Lietuvių
- Magyar
- Na Vosa Vakaviti
- 日本語
- Polski
- Русский
- Suomi
- Svenska
- தமிழ்
- Тоҷикӣ
- Tiếng Việt
- 中文

thank you

1. An expression of gratitude or politeness, in response to something done or given.

Synonyms [edit]

- cheers (*informal*), thanks, thanks very much, thank you very much, thanks a lot, ta (*UK*, *Australia*), thanks a bunch (*informal*), thanks a million (*informal*), much obliged, gracias, muchos gracias, mooch ass grassy ass

Translations [edit]

- Abkhaz: иҭабуп (ab) (iţ°abup)
- Adangme: mo tsumi
- Afar: gadda ge
- Afrikaans: dankie (af), baie dankie (af)
- Albanian: faleminderit (sq), ju falem nderit (sq)
  - Gheg: falimineres
- Aleut: qaĝaasakuq (Atkan), qaĝaalakux̂ (Eastern)
- Alutiiq: quyanaa
- American Sign Language: OpenB@Chin-PalmBack OpenB@FromChin-PalmUp
- Amharic: አመሰግናለሁ (am) (amesegenallo)
- Arabic: شُكْرًا (ar) (šúkran)
- Aramaic: ܬܘܕܝ (tawdi)
- Armenian: շնորհակալություն (hy) (šnorhakalut'yun), շնորհակալ եմ (hy) (šnorhakal em), (*colloquially*) մերսի (hy) (mersi)
- Kiowa: áho
- Kongo: ntondele
- Korean: 고맙습니다 (ko) (gomapseumnida), 감사합니다 (ko) (gamsahamnida), 고마워 (ko) (gomawoe)
- Kurdish: spas (ku), سوپاس (ku)
- Kyrgyz: рахмат (ky) (rahmat), ыракмат (ky) (irakmat)
- Ladin: giulan
- Lao: ຂອບໃຈ (lo) (kööp cai)
- Latin: benignē dīcis (la), tibi gratiās agō (la), (*informal*) gratiās (la) ***f pl***, gratiās agō (la)
- Latvian: paldies (lv)
- Lithuanian: ačiū (lt)
- Luo: erokamano
- Macedonian: благодарам (mk) (blagódaram) (*formal*), фала (mk) (fála) (*informal*)
- Malagasy: misaotra (mg)