CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

Introduction to Computational Terminology

Alberto Simões and José João Almeida

ambs@ilch.uminho.pt, jj@di.uminho.pt

Summer School of Linguistics
10th September, 2013

Outline

Outline

terminology
noun (plural *terminologies*)

the body of terms used with a particular technical application in a subject of study, theory, profession, etc.: *the terminology of semiotics*

Oxford Dictionaries

Terminology

Main points:

- A collection of terms;
- Terms from a specific domain;

Terminology

Main points:

- A collection of terms;
- Terms from a specific domain;

Problems:

Q. **What is a term?**

Q. **What is specific from a domain?**

Terminology

Main points:

- A collection of terms;
- Terms from a specific domain;

Problems:

Q. **What is a term?**

A. Some sequence of words, usually a noun phrase. Definition not relevant, as we will only find candidates (human has the final word);

Q. **What is specific from a domain?**

Terminology

Main points:

- A collection of terms;
- Terms from a specific domain;

Problems:

Q. **What is a term?**

A. Some sequence of words, usually a noun phrase. Definition not relevant, as we will only find candidates (human has the final word);

Q. **What is specific from a domain?**

A. Has a specific meaning in that domain, or refers to entities specific from that domain;

Computational Terminology

The use of computational techniques to help discovering the relevant terms for a specific domain.

Alberto Simões

Outline

Corpora for Terminology Building

- The use of a specific domain text or texts in order to understand what is that domain terminology is relevant;
- Words in context give more information than alone;
- There is no automatic method to extract specific domain terminology from a specific domain corpus;
- Nevertheless, there are automatic method to obtain candidate terms, that can later be analysed and incorporated in a terminology, or just discarded.

Outline

Words n-Grams

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech.

The items in question can be phonemes, syllables, letters, words or base pairs according to the application. n-grams are collected automatically from a text or speech corpus.

One-Grams

1-Grams are usually known as words/tokens. :-)

Peter Piper picked a peck of pickled peppers.
A peck of pickled peppers Peter Piper picked.
If Peter Piper picked a peck of pickled peppers,
Where's the peck of pickled peppers Peter Piper picked?

peter	4
piper	4
picked	4
a	2
peck	4
of	4
pickled	5
⋮	⋮

Bigrams

All sequences of two words/tokens found in the text.

Peter Piper picked a peck of pickled peppers.
A peck of pickled peppers Peter Piper picked.
If Peter Piper picked a peck of pickled peppers,
Where's the peck of pickled peppers Peter Piper picked?

peter piper	4
piper picked	4
picked a	2
a peck	3
peck of	4
of pickled	4
pickled peppers	4
⋮	⋮

Top occurring trigrams for a real corpus

CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

in accordance with	31148	of this directive	6679
referred to in	27581	a member state	6306
the member states	16999	on the basis	6292
accordance with the	16535	the european parliament	6274
of the european	14772	the basis of	6265
laid down in	13301	and in particular	6225
to in article	13211	down in article	6200
having regard to	12588	of the community	5958
regard to the	11416	accordance with article	5758
member states shall	11392	to in paragraph	5690
in order to	10563	opinion of the	5599
in the case	10029	the opinion of	5191
the provisions of	9825	the competent authorities	5074
the case of	9575	for the purposes	5024
provided for in	9560	the purposes of	4946
the member state	9360	with the procedure	4878
of the member	8656	to the commission	4843
the commission shall	8013	the european community	4834

n-grams frequency

- n-Grams are usually computed together with their occurrence count — or frequency;
- In some situations, like statistic language models, other type of measures are also computed (probability — relative frequency; conditional probability, etc);
- One-grams frequency doesn't help much on term candidate extraction — they just say that a word is more or less frequent.
- *n*-grams for $n \geq 2$ can help finding sequence of words that occur lot of times.

Stop Words and Lexical Difference

- There are words that rarely occur in terminology;
- At least, they rarely occur in the beginning or end of a multi-word term;
- For example, pronouns, articles, prepositions;
- These words are usually known as stop words;
- It is easy to find bigger or smaller lists of stop words for every language;
- We can ignore these words when computing *n*-grams.

Detecting stop-words

in accordance with	31148
referred to in	27581
the member states	16999
accordance with the	16535
of the european	14772
laid down in	13301
to in article	13211
having regard to	12588
regard to the	11416
member states shall	11392
in order to	10563
in the case	10029
the provisions of	9825
the case of	9575
provided for in	9560
the member state	9360
of the member	8656
the commission shall	8013

of thisi directive	6679
a member state	6306
on the basis	6292
the european parliament	6274
the basis of	6265
and in particular	6225
down in article	6200
of the community	5958
accordance with article	5758
to in paragraph	5690
opinion of the	5599
the opinion of	5191
the competent authorities	5074
for the purposes	5024
the purposes of	4946
with the procedure	4878
to the commission	4843
the european community	4834

Replacing stop words by a special token

CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

<tk> member states	32517	competent authority <tk>	3507
member states <tk>	30108	annex ii <tk>	3429
<tk> member state	19345	commission regulation <tk>	3171
member state <tk>	17882	<tk> commission regulation	2967
council directive <tk>	7869	commission decision <tk>	2545
<tk> council directive	7129	<tk> customs authorities	2542
<tk> european parliament	5397	<tk> commission decision	2429
council regulation <tk>	5259	customs authorities <tk>	2410
european parliament <tk>	5125	<tk> european economic	2285
<tk> council regulation	4995	<tk> administrative provisions	2017
<tk> competent authorities	4964	<tk> contracting parties	2010
competent authorities <tk>	4736	conditions laid <tk>	1998
procedure laid <tk>	4472	contracting parties <tk>	1779
<tk> treaty establishing	4375	commission directive <tk>	1764
treaty establishing <tk>	4373	detailed rules <tk>	1738
<tk> competent authority	3694	<tk> community industry	1728
official journal <tk>	3530	<tk> contracting party	1702

Trigrams that doesn't include stop words

member states relating	1523	regional economic integration	263
member state concerned	1200	median longitudinal plane	258
veterinary medicinal products	955	plant protection product	249
maximum residue limits	814	separate technical unit	246
physically modified derivatives	700	national regulatory authorities	241
european economic community	691	apply mutatis mutandis	241
community trade mark	538	common technical regulation	229
member states concerned	508	separate technical units	226
plant protection products	464	emission limit values	219
home member state	442	technically permissible maximum	215
host member state	388	maximum residue levels	212
council common position	377	retail trade services	200
community plant variety	368	temporary importation procedure	196
european atomic energy	346	medicinal products intended	195
animal health conditions	342	community transit procedure	195
authorised representative established	327	atomic energy community	193
implementing powers conferred	311	classical swine fever	189

Basic Lexical Difference

- What if we remove not just stop words, but common words?
- It is not that usual to find Osteoarthritis in common text. Therefore, it should be some kind of a domain term.
- We can obtain a list of common words from a generic corpus (say, jornalistic text) and subtract that lexicon from the one-grams we obtained.
- Result should include good term candidates!

Basic Lexical Difference - Bad Experiment

- Two random abstracts from PubMed articles related with cirrhosis;
- Top 1 000 occurring words in English;
- Compute one-grams on the abstracts;
- Subtract the top occurring words.

Before		After	
liver	8	liver	8
is	7	myofibroblast	6
fibrosis	6	fibrosis	6
myofibroblast	6	pathway	5
pathway	5	kidney	5
kidney	5	interstitial	4
expression	5	β-catenin	3
interstitial	4	target	3
signaling	3	signaling	3
target	3	genes	3
differentiation	3	differentiation	3
diseases	3	medullary	3
medullary	3	renal	3
antioxidant	3	adult	3

Lexical Distribution Difference

- Previous example could benefit a bigger standard lexicon list;
- Abstracts are crowded with terminology, and few other words;
- Long lists may include words than are considered terminology! Example, for Informatics, folder or file can be terms.

Lexical Distribution Difference

- Previous example could benefit a bigger standard lexicon list;
- Abstracts are crowded with terminology, and few other words;
- Long lists may include words than are considered terminology! Example, for Informatics, folder or file can be terms.

- Instead of considering words as present or not, use their frequency;
- For instance, compute relative frequency and compare/subtract;
- Use a distribution comparison metric;
  - ex., Kullback-Leibler terms: $\log\left(\frac{P(i)}{Q(i)}\right) P(i)$

Outline

Pointwise Mutual Information

The Mutual Information (MI) is a quantity that measures the mutual dependence of two random variables X and Y.

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces uncertainty about the other.

Pointwise Mutual Information

When computing Mutual Information for two specific outcomes, the Pointwise Mutual Information (PMI) let us measure their mutual dependence:

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Given the number of tokens in the document N, and the number of occurrences for x, $Oc(x)$: $P(x) = \frac{Oc(x)}{N}$
- Given the number of tokens in the document N, and the number of occurrences for bigram x, y, $Oc(x, y)$: $P(x, y) = \frac{Oc(x,y)}{N}$

Pointwise Mutual Information

CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

Sorted by occurrence count

sonic fabric	14	7.3566
black holes	9	8.0912
black hole	7	8.0912
cassette tape	6	8.4968
build things	4	9.5348
smartphone makers	3	9.0087
alyce santoro	3	8.0912
like scratching	3	9.0087
barnard said	3	8.3042
milky way	3	9.1787
possible black	3	7.6762
neutron star	3	8.8567
just right	3	8.5937
records backwards	3	10.5937

Sorted by PMI

special shuttle	1	12.1787
immediately reminded	1	12.1787
remain aware	1	12.1787
richard branson	1	12.1787
supercooled pods	1	12.1787
richie havens	1	12.1787
auspicious locations	1	12.1787
jimi hendrix	1	12.1787
account settings	1	12.1787
baggage carousel	1	12.1787
buddhist prayer	1	12.1787
reinvents electronics	1	12.1787
melbourne institute	1	12.1787
cow manure	1	12.1787

From a very small corpus constructed with 5 CNN news stories.

Outline

Morphology Patterns

- Commonly, terms are nouns or noun phrases;
- Sometimes some verbs are also interesting;
- Typically the morphological structure of terms is well known;
- There is software that compute morphological information about each word in a sentence;
- We can use that information to obtain better term candidates. specify terms part-of-speech, genre, number, verb tenses, etc. . .

Morphological Analysis

How it (usually) works:

1. A tokenizer and a splitter split sentences into tokens and sentences;
   (different tools use them in different order, some as a single tool)
2. A morphological analyzer associates possible analysis to each word;
   (does not cope with ambiguity, just tags all possible analysis)
3. A Tagger or Parser choose the more likely analysis;
   (uses knowledge from manual annotated corpora, and machine learning algorithms)

Morphological Patterns - Examples

CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

	Noun Noun Noun
659	Community trade mark
483	plant protection products
475	EEC component type-approval
448	document number C
320	Community transit procedure
290	plant protection product
288	Community plant variety
257	EC type-examination certificate
214	EC component type-approval
176	EEC pattern approval
157	African swine fever
155	three-wheel motor vehicles
155	foot-and-mouth disease virus
153	conformity assessment procedures
148	emission limit values

	Adjective Adjective Noun
912	veterinary medicinal products
453	common agricultural policy
365	separate technical unit
291	separate technical units
265	median longitudinal plane
223	regional economic integration
202	competent national authorities
200	trans-European high-speed rail
199	sound financial management
189	veterinary medicinal product
182	certain agricultural products
176	national regulatory authorities
175	common technical regulation
168	certain third countries
166	other third countries
166	definitive anti-dumping duty
162	certain dangerous substances

CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

Term-o-Matic

http://termomatic.com

Outline

Sentence Alignment

Sentence alignment is the task of detecting translation relationships between sentences in parallel corpora.

If s_α is a sentence in a language $\mathcal{L}_\alpha$ and s_β is a sentence in a language $\mathcal{L}_\beta$, the alignment process creates the pair (s_α, s_β) if (there is a high probability that) s_β is a translation of s_α.

We will use the term Translation Unit to refer to a sentence pair.

Word Alignment

The Word Alignment is the task of detecting translation relationships between words or terms in sentence-aligned parallel corpora.

Word Alignment

The Word Alignment is the task of detecting translation relationships between words or terms in sentence-aligned parallel corpora.

There are two trends on word alignment:

- for each translation unit, create a link between every word and its translation;
- for the complete corpora, obtain a relationship between a word and a set of probable translations, together with a confidence measure (a kind of translation probability);

- Obtained with one of the word alignment methods;
- Define a relationship between a word and a set of probable translations;

$$\mathcal{T}(\text{europe}) = \begin{cases} \text{europa} & 94.7\% \\ \text{europeus} & 3.4\% \\ \text{europeu} & 0.8\% \\ \text{europeia} & 0.1\% \end{cases}$$

$$\mathcal{T}(\text{stupid}) = \begin{cases} \text{estúpido} & 47.6\% \\ \text{estúpida} & 11.0\% \\ \text{estúpidos} & 7.4\% \\ \text{avisada} & 5.6\% \\ \text{direita} & 5.6\% \\ \text{impasse} & 4.5\% \\ \text{ocupado} & 3.8\% \end{cases}$$

Outline

Monolingual Techniques

- Same techniques used for Monolingual corpora can be used for parallel corpora;
- Detect terms in languages, separately, and align by features:
  - occur in the same translation unit;
  - has similar frequency counts;
  - use translation dictionary information;

Outline

Translation Matrix

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

- Using the probabilistic translation dictionaries we are able to construct a translation matrix;
- Each cell has a translation probability obtained from the dictionary;

Translation Patterns

- Translation changes word order (for some language pairs!);
- This change can be foreseen;
- This change can be defined formally as a pattern;
- These patterns can be used to obtain term candidates.

Translation Pattern 1: ABBA

	Jogos	Olímpicos
Olimpic		X
Games	X	

Formally,

$$\mathcal{T}(A \cdot B) = \mathcal{T}(B) \cdot \mathcal{T}(A)$$

Or in the tool syntax:

```
[ABBA] A B = B A
```

Translation Pattern 2: IDH

	índice	de	desenvolvimento	humano
human				X
development			X	
index	X			

$$\mathcal{T}(I \cdot \text{"}de\text{"} \cdot D \cdot H) = \mathcal{T}(H) \cdot \mathcal{T}(D) \cdot \mathcal{T}(I)$$

```
[IDH] I "de" D H = H D I
```

	protocolo	de	transferência	de	ficheiros
file					X
transfer			X		
protocol	X				

$$\mathcal{T}(P \cdot \text{"}de\text{"} \cdot T \cdot \text{"}de\text{"} \cdot F) = \mathcal{T}(F) \cdot \mathcal{T}(T) \cdot \mathcal{T}(P)$$

```
[FTP] P "de" T "de" F = F T P
```

Patterns in Translation Matrix

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

The two boxes correspond to the following two patterns:

- `[P1] F "de" N A = A F "of" N`
- `[P2] A B C = C B A`

Terms extracted using A B = B A

21007	união europeia	⇒	european union
9301	parlamento europeu	⇒	european parliament
4171	direitos humanos	⇒	human rights
3504	estados unidos	⇒	united states
2353	mercado interno	⇒	internal market
1911	posição comum	⇒	common position
1826	países candidatos	⇒	candidate countries
1776	comissão europeia	⇒	european commission
1708	conselho europeu	⇒	european council
1629	saúde pública	⇒	public health
1558	direitos fundamentais	⇒	fundamental rights
1546	nações unidas	⇒	united nations
1337	países terceiros	⇒	third countries
1294	conferência intergovernamental	⇒	intergovernmental conference
1258	fundos estruturais	⇒	structural funds

Terms extracted using A "de" B = B A

729	plano de acção	⇒	action plan
722	conselho de segurança	⇒	security council
680	processo de paz	⇒	peace process
582	mercado de trabalho	⇒	labour market
580	pena de morte	⇒	death penalty
492	pacto de estabilidade	⇒	stability pact
431	política de defesa	⇒	defence policy
353	acordo de associação	⇒	association agreement
348	protocolo de quioto	⇒	kyoto protocol
343	programa de acção	⇒	action programme
259	branqueamento de capitais	⇒	money laundering
258	comité de conciliação	⇒	conciliation committee
241	política de concorrência	⇒	competition policy
226	processo de conciliação	⇒	conciliation procedure
217	requerentes de asilo	⇒	asylum seekers

Terms extracted using A B C = C B A

531	política agrícola comum	⇒	common agricultural policy
418	banco central europeu	⇒	european central bank
329	tribunal penal internacional	⇒	international criminal court
166	aliança livre europeia	⇒	european free alliance
156	modelo social europeu	⇒	european social model
153	partidos políticos europeus	⇒	european political parties
83	fundo monetário internacional	⇒	international monetary fund
75	política externa comum	⇒	common foreign policy
66	organização marítima internacional	⇒	international maritime organisation
65	própria união europeia	⇒	european union itself
65	fundo social europeu	⇒	european social fund
55	direitos humanos fundamentais	⇒	fundamental human rights
45	relações económicas externas	⇒	external economic relations
45	homens e mulheres	⇒	women and men
45	agência espacial europeia	⇒	european space agency

Terms extracted: I "de" D H = H D I

95	mandato de captura europeu	⇒	european arrest warrant
85	fontes de energia renováveis	⇒	renewable energy sources
80	mandado de captura europeu	⇒	european arrest warrant
67	sistemas de segurança social	⇒	social security systems
64	zona de comércio livre	⇒	free trade area
55	força de reacção rápida	⇒	rapid reaction force
54	orientações de política económica	⇒	economic policy guidelines
46	planos de acção nacionais	⇒	national action plans
46	direitos de propriedade intelectual	⇒	intellectual property rights
33	sistema de alerta rápido	⇒	rapid alert system
29	política de defesa comum	⇒	common defence policy
29	método de coordenação aberta	⇒	open coordination method
27	método de coordenação aberto	⇒	open coordination method
27	conselho de empresa europeu	⇒	european works council
25	acordo de comércio livre	⇒	free trade agreement

Adding Morphological Constraints

- The pattern language supports constraints;
- Constrains can be of different types;
- The most interesting are the morphological ones:

```
[ABBA] A B[CAT<-adj] = B[CAT<-adj] A
```

- With this kind of constrain we can force the words in specific positions to be of specific morphological category.

Outline

Monolingual Terms Translation

- Proposed by Gornostay et al, 2012;
- Extract terminology in one language using morphology patterns;
- Translate to target language using a translation dictionary, obtaining all possible translations;
- Search obtained translations in the target language;

Outline

Further Reading

- Church, Kenneth Ward and Patrick Hanks. 1990. Word association Norms, mutual information, and lexicography. In Computational Linguistics, 16(1), March, 22–29.
- Pantel, Patrick and Dekang Lin. 2001. A Statistical Corpus-Based Term Extractor. In Advances in Artificial Intelligence, Springer, 36–46.
- Guinovart, Xavier Gomez and Alberto Simões. 2009. Parallel corpus-based bilingual terminology extraction. In 8th International Conference on Terminology and Artificial Intelligence, Toulouse, France, November, 18–20.
- Gornostay, Tatiana, Anita Ramm, Ulrich Heid, Emmanuel Morin, Rima Harastani, Emmanuel Planas. 2012. Terminology Extraction from Comparable Corpora for Latvian. In Baltic HLT. 66–73.