

DEXTER: Automatic term extraction of domain-specific glossaries for language teaching

Carlos Periñán
Eva M. Mestre

Paper structure

1. Introduction: ATE
2. DEXTER:
 - 2.1. SRC: A metric for term extraction
 - 2.1.1. Termhood
 - 2.1.2. Unithood
3. Dictionaries and thesauri as reference lists
4. Experimentation
5. Discussion of results
6. Conclusions

Approaches to Automatic Term Extraction (ATE)

Linguistic:

linguistic structures

- POS tagging
- phrase chunking

Statistic:

based exclusively on
frequency

Hybrid:

most productive and
widely used

- linguistic
- statistical

DEXTER

ATE Workbench

Online access

Specialized corpus

Analysis (hybrid) :

(1) [LINGUISTIC] Shallow lexical filters

(1) Stopword lists

(2) No POS tagging

(2) [STATISTIC]

Parameterized Composite Metric

Metric:



- Composite
- User-adjustable
- Unit of analysis: **stemmed n-gram**

Parameters:

- user adjustable

METRICS: What do they tell us?

unithood:

Stability in the relationship within the members of an **n-gram**

- an adaptation of Park, Byrd and Boguraev's Term Cohesion (2002)

termhood:

To what extent an **n-gram** describes/represents the specialized domain

- TF-IDF (Salton, Wong, and Yang 1975; Salton, Yang, and Yu 1975)
- The weight of a term is determined by the relative frequency of the term in a certain document

METRICS

unithood: Stability in the relationship of an **n-gram**

termhood: an **n-gram** represents the specialized domain

Parameters in **DEXTER**

Salience

Term in relation to the document:

Unique words, keywords

Relevance

Term in relation to the domain:

Compared to a reference, general corpus

Cohesion

Term in relation to other terms within an n-gram

Not only proportional to the frequency of the n-gram, but frequency of items which compose it alone

Salience

Relevance

Cohesion

SALIENCE: Unique words, keywords

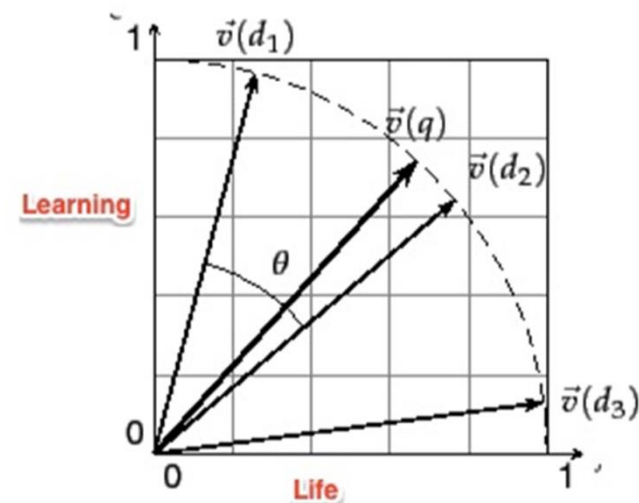
$$TF-IDF(g) = TF(g) * IDF(g) * NORM(g)$$

TF: Term Frequency

IDF: Inverse Document Frequency

NORM: Normalization factor

Vector space models



S alience

R elevance

C ohesion

RELEVANCE: Term in the domain

specialized corpus / reference corpus

- Adaptation of **Weirdness** (Ahmad, Gillam & Tostevin, 2000)

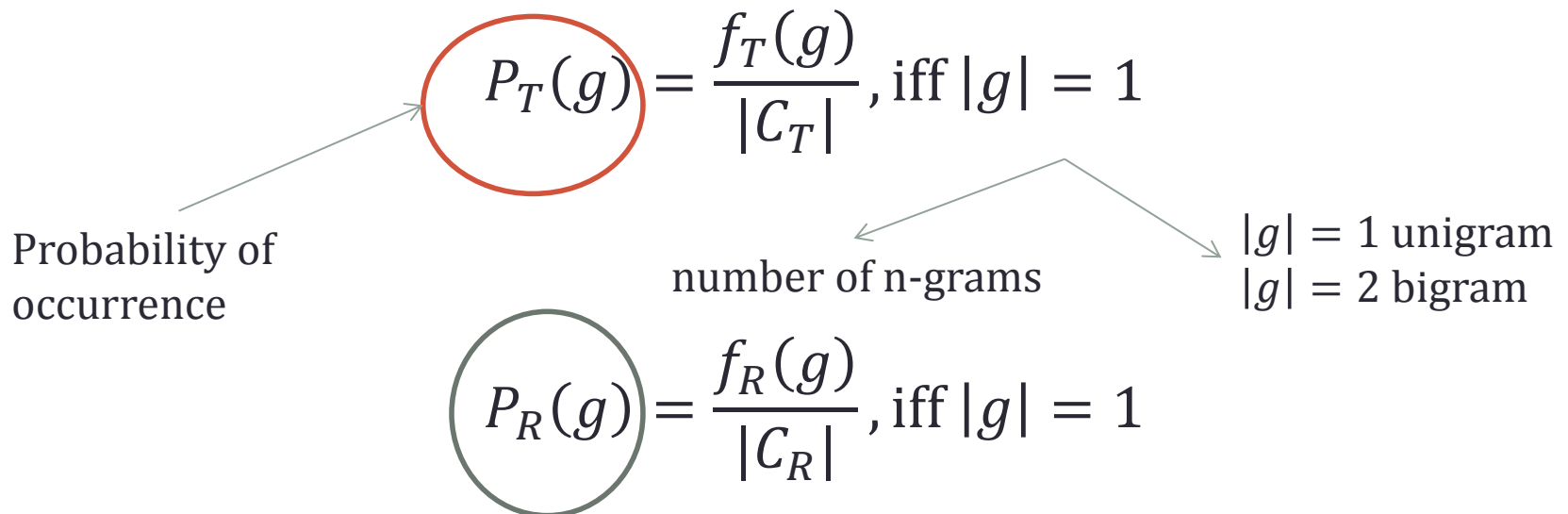
$$R_T(g) = 1 - \frac{1}{\log_2 \left(2 + \frac{P_T(g)}{P_R(g)} \right)}$$

DEXTER

Normalization factor (0-1)

Salience**R**elevance**C**ohesion

Adaptation of Weirdness



S alience

R elevance

C ohesion

COHESION: consistency of the relation in an n-gram.

Adaptation of Park, Byrd et al.

$$C_T(g) = 1 - \frac{1}{\log_2 \left(2 + \frac{f_T(g)}{\sqrt{\prod_{k_i \in g} f_T(k_i)}} \times F \right)}, \text{ iff } |g| > 1$$

DEXTER

Mean: arithmetic and not geometric.

Normalization factor

$$F = \begin{cases} 1, & \text{iff } f_T(g) = 1 \\ \log_2(f_T(g)), & \text{iff } f_T(g) > 1 \end{cases}$$

Normalization factor NOT for terms with frequency=1

COMPOSITE MEASURE SRC

$$SRC_T(g) = termhood(g) + unithood(g)$$

$$termhood(g) = S_T(g) * \alpha + R_T(g) * \beta$$

$$unithood(g) = \begin{cases} 0, & \text{iff } |g| = 1 \\ C_T(g) * \gamma, & \text{iff } |g| > 1 \end{cases}$$

Working with terminology: dictionaries as reference lists

Specialised glossaries, dictionaries and thesauri

- classificatory structures of knowledge organisation
- common terminological tools used in scientific and technical documentation
- Also used for **pedagogical purposes**
 - LSP courses: lexis in specific domains
 - domain-specific glossaries

Working with terminology: dictionaries as reference lists

- **Lexical resources as reference lists in ATE validation**
 - gold standards in terminology: exemplars of quality /not of perfection.
 - pace at which science and technology require coinage of new terms.
 - neologisms: studied, agreed upon, accepted and finally incorporated into the language, Cabré (2007).
 - lexical resources take a snapshot of language in use at the time of compilation.
 - ATE systems usually extract terms not present in reference lists.

THUS, reference lists cannot be solely relied upon.

- For a multilingual platform such as DEXTER: best option of gold standard **IATE**, the multilingual term database of the European Union.

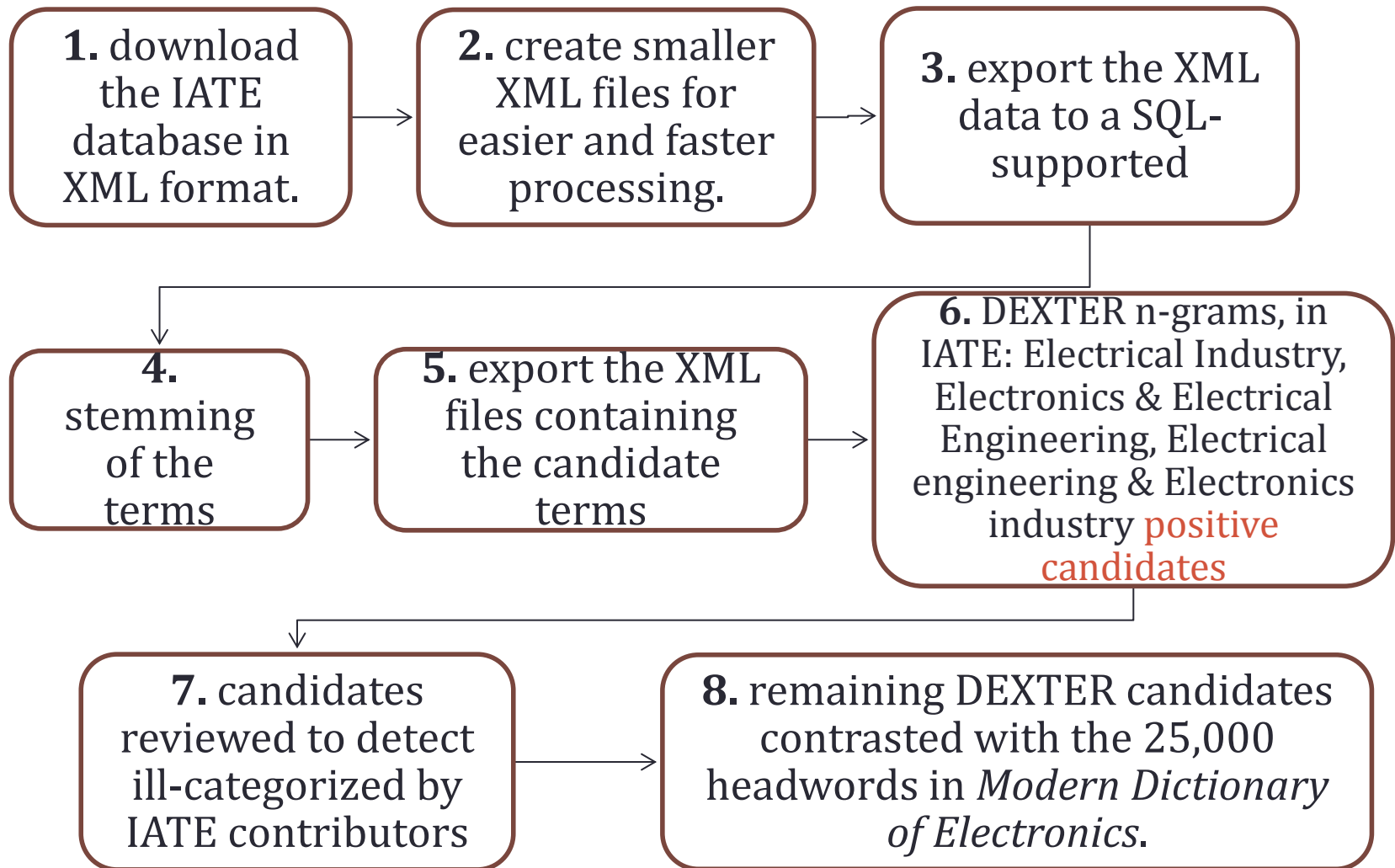
EXPERIMENTATION

- Small corpus of 46 documents (78,988 tokens)
- Topic: electronics.
- <http://www.electronics-tutorials.ws>

Subtopic	Documents
AC theory	17
DC circuits	10
Input-output devices	8
Miscellaneous circuits	11

- DEXTER extracted 1,268 **unigrams**, 1,143 **bigrams** and 377 **trigrams**.
- most ngrams: **nouns** / **noun phrases**,
also **adjectives** (e.g. *capacitive, harmonic, inductive, or sinusoidal*)
and a few **verbs** (e.g. *amplify or regulate*).

Hybrid procedure of evaluation for SRC. Type-A EVALUATION



Hybrid procedure of evaluation for SRC. Type-B EVALUATION

- Manual validation of candidates not selected as terms in the previous stage.
- Applied to the outcome of the type-A evaluation, so that the problem of subjectivity was minimized.

Hybrid procedure of evaluation for SRC. Type-B EVALUATION

- Corollary 1. **Several lexical items** ($item_i, item_{i+1}, \dots, item_n$) can be treated as a **single lexicalization** ($complexItem$) as long as the definition of $complexItem$ can be found in the corpus or in any other document resource.
- Corollary 2. **The definition of $complexItem$** must go **beyond** the simple combination of the meanings of $item_i, item_{i+1}, \dots, item_n$.
- Corollary 3. **The definition of $complexItem$** must **contain at least one lexical item**, different from $item_i, item_{i+1}, \dots, item_n$, which is typically **used to describe the corpus domain**.

The type-B evaluation revealed 12 unigrams, 38 bigrams and 64 trigrams as false negative terms among the top 200 SRC-ranked ngrams

Results for unigrams

#candidates	S	R	f
1-40	0.82	0.45	0.75
41-80	0.57	0.65	0.37
81-120	0.45	0.70	0.42
121-160	0.45	0.62	0.27
161-200	0.47	0.55	0.50
	0.55	0.59	0.46

Precision in the Type-A evaluation of unigrams.

#candidates	SRC	S	R	f
1-40	0.82	0.85	0.60	0.75
41-80	0.72	0.57	0.75	0.37
81-120	0.67	0.45	0.70	0.42
121-160	0.65	0.47	0.65	0.27
161-200	0.57	0.47	0.57	0.50
	0.69	0.56	0.65	0.46

Precision in the Type-B evaluation of unigrams.

Discussion of results

- The integration of two large gold-standard data sets, (IATE / electronics dictionary), is not sufficient.
- Manual validation increased the number of recognized terms proportionally to the complexity of the ngram

	S	R	C
Unigrams	+0.01	+0.06	
Bigrams	+0.16	+0.14	+0.12
Trigrams	+0.24	+0.31	+0.30

- Best precision: SRC for top 200 unigrams, bigrams and trigrams:
 - the combination of metrics can improve the performance of this ATE system.
 - the distribution of terms is better with SRC than with the single metrics.
- For ATE systems focusing on small- and medium-sized specialized corpora, SRC is an efficient, precise measure.

Discussion of results

Categories of false positive candidates among the top 200 SRC-ranked ngrams

- i. **Common words**, such as *average*, *maximum* or *value*.
- ii. Words typically used in the **description of other specialized domains**, e.g. *complex number*, *root mean square*.
- iii. **Words nested in multi-word terms** of the given domain:
 - unigrams – *nodal* (*nodal voltage analysis*),
 - bigrams – *width modulation* (*pulse width modulation*),
 - trigrams – *permanent magnet DC* (*permanent magnet DC motor*)
 - proper nouns are almost always part of complex terms: *Kirchoffs* (*Kirchoffs Circuit Law*, *Kirchoffs Current Law* and *Kirchoffs Voltage Law*),
- iv. **Mathematical variables and symbols**: Variables X_L , X_C and X_T were converted into the ngrams XL, XC y XT:

Conclusions

SRC

composite measure for term extraction

user-adjustable metric based on

- salience
- relevance
- cohesion

DEXTER

multilingual platform for

- data mining
- terminology management

small- medium-sized specialized corpora.

dictionaries & thesauri: insufficient to evaluate candidate terms

semi-automatic hybrid approach with human validation

DEXTER: Automatic term extraction of domain-specific glossaries for language teaching

Carlos Periñán
Eva M. Mestre