Extraction of Folksonomies from Noisy Texts

Wim De Smet Marie-Francine Moens

LIIR
Departement of Computer Science
K.U.Leuven
Belgium

September 3, 2007

Project $A4MC^3$
Architectures for Mobile Community Content Creation

Goal

Automatic creation of folksonomy in a geographic community
Input: texts from amateur authors

Definition

Folksonomy: taxonomy created by end-users of the system

Goal

Automatic creation of folksonomy in a geographic community
Input: texts from amateur authors

Definition

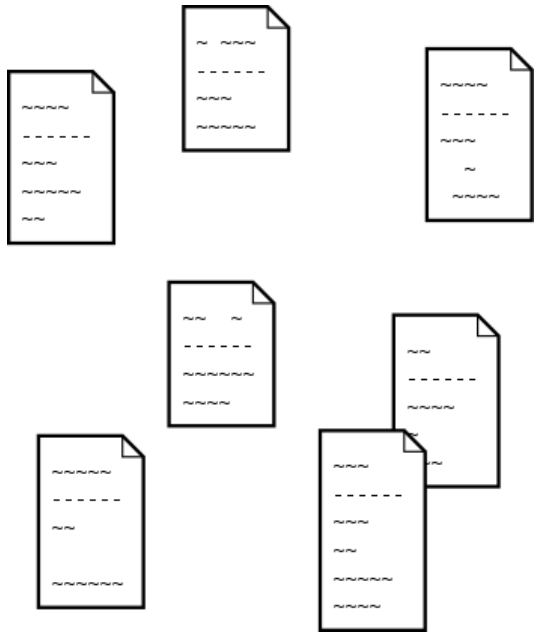
Folksonomy: taxonomy created by end-users of the system

Goal

Automatic creation of folksonomy in a geographic community
Input: texts from amateur authors

Definition

Folksonomy: taxonomy created by end-users of the system

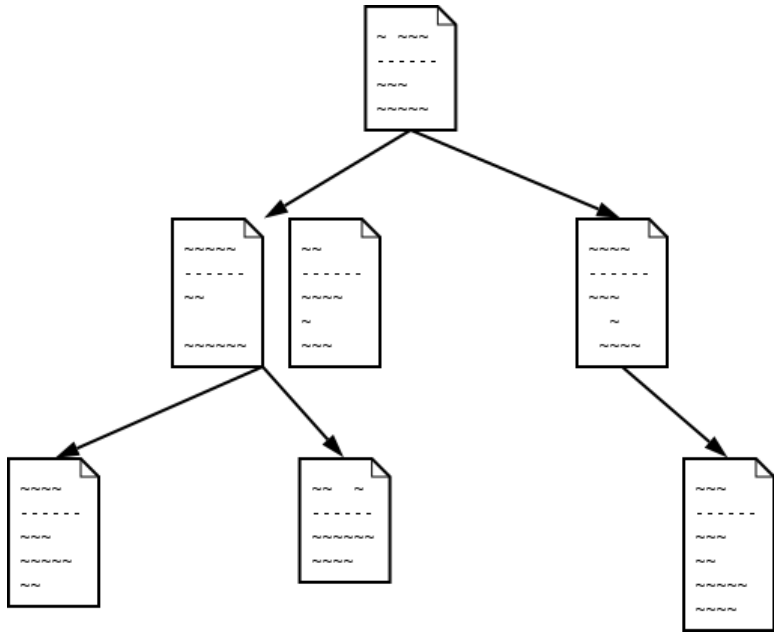


Goal

Automatic creation of folksonomy in a geographic community
Input: texts from amateur authors

Definition

Folksonomy: taxonomy created by end-users of the system



Problem 1: Users

- Non-professional authors $\Rightarrow$ spelling errors
- Geographic community $\Rightarrow$ dialect language

Consequences:

- unknown words
- no concensus on spelling
- ...

Solution:

- correct spelling errors
- resolve dialect words to standard words

Problem 1: Users

- Non-professional authors $\Rightarrow$ spelling errors
- Geographic community $\Rightarrow$ dialect language

Consequences:

- unknown words
- no concensus on spelling
- ...

Solution:

- correct spelling errors
- resolve dialect words to standard words

Problem 1: Users

- Non-professional authors $\Rightarrow$ spelling errors
- Geographic community $\Rightarrow$ dialect language

Consequences:

- unknown words
- no concensus on spelling
- ...

Solution:

- correct spelling errors
- resolve dialect words to standard words

Problem 2: Language

Used language is Dutch

- No common parsers available
- Small support on dictionaries
- ...

Outline

Definitions

Definition

Dialect: Geographical variation of language

Example

	Flemish	East-Flemish West-Flemish
Flemish Region	Brabant	Brabantish Antwerp
	Limburg	Brussels Limburg Kempen

Definition

Standard Language: Official dialect

Definitions

Definition

Dialect: Geographical variation of language

Example

	Flemish	East-Flemish West-Flemish
Flemish Region	Brabant	Brabantish Antwerp
	Limburg	Brussels Limburg Kempen

Definition

Standard Language: Official dialect

Definitions

Definition

Dialect: Geographical variation of language

Example

	Flemish	East-Flemish West-Flemish
Flemish Region	Brabant	Brabantish Antwerp
	Limburg	Brussels Limburg Kempen

Definition

Standard Language: Official dialect

Differences between dialects

- grammar
- vocabulary $\Rightarrow$ unknown words
- phonology $\Rightarrow$ different spelling

Example

grasmaaier $\Rightarrow$ graasmesjien

Example

bananenschil $\Rightarrow$ benanesjèl

Differences between dialects

- ~~grammar~~
- vocabulary $\Rightarrow$ unknown words
- phonology $\Rightarrow$ different spelling

Example

grasmaaier $\Rightarrow$ graasmesjien

Example

bananenschil $\Rightarrow$ benanesjèl

Differences between dialects

- ~~grammar~~
- vocabulary ⇒ unknown words
- phonology ⇒ different spelling

Example

grasmaaier ⇒ graasmesjien

Example

bananenschil ⇒ benanesjèl

Differences between dialects

- ~~grammar~~
- vocabulary ⇒ unknown words
- phonology ⇒ different spelling

Example

grasmaaier ⇒ graasmesjien

Example

bananenschil ⇒ benanesjèl

Summarization of possible errors

Errors to distinguish:

- Spelling error in standard word
- Phonological variant of standard word
- New unknown dialect words

Summarization of possible errors

Errors to distinguish:

- Spelling error in standard word
- Phonological variant of standard word
- New unknown dialect words

Summarization of possible errors

Errors to distinguish:

- Spelling error in standard word
- Phonological variant of standard word
- New unknown dialect words

Terms & Notation

cf(*term*): Corpus frequency of term
number of times term occurs in corpus

ded: Dialect Edit Distance: measure of similarity of dialect and standard words

D: Dictionary of correctly spelled words

L: Lexicon of words apparent in corpus

Our goal: Update *D* and correct *L* while scanning each text, using *cf* and *ded* of each term in *L*.

Algorithm

$threshold_{cf}$: min corpus frequency
$threshold_{ded}$: min dialect edit distance
if $(t \in D)$ **then**
  return t
end if
if $(cf(t) > threshold_{cf})$ **then**
  $D.add(t)$ {frequency assumption}
  return t
else
  select n words from L with smallest ded to t, $n \geq 1$
  if (min ded from D) $<$ $threshold_{ded}$ **then**
    return n words
  else
    $D.add(t)$ {edit distance assumption}
    return t
  end if
end if

Edit Distance

Definition

Edit Distance: Measure on similarity of words

word vs. bored

		b	o	r	e	d
	0	1	2	3	4	5
w	1	1	2	3	4	5
o	2	2	1	2	3	4
r	3	3	2	1	2	3
d	4	4	3	2	2	2

Edit Distance

Definition

Edit Distance: Measure on similarity of words

word vs. bored

		b	o	r	e	d
	0	1	2	3	4	5
w	1	1	2	3	4	5
o	2	2	1	2	3	4
r	3	3	2	1	2	3
d	4	4	3	2	2	2

```
wor d
s..i.
bored
```

Learning adaptive costs

Adaptive costs: low for operations, typical for dialect
Use dictionary of standard-dialect word pair
For each pair: calculate rules = operations+surrounding context

Example

```
aafb eje
.d..i.s.i
a fbieden
Second rule: Insert i, context=fb.ej
```

Learning adaptive costs, cont.

Cost rule dependant on:

- number of times operation appears within (subset of) context in dictionary (O)
- total number of times context appears in dictionary (C)

Cost formula takes into account:

- ratio $\frac{O}{C}$
- C

Definition

$$f = g \cdot \frac{O^2}{C^2} - 2 \cdot g \cdot \frac{O}{C} + 1$$
$$\textit{where} \quad g = 1 - \frac{1}{1+log(1+C)}$$

Applying adaptive operation costs

Apply standard dialect edit distance algorithm.
For each operation: find rule containing operation and (subset of) context with lowest cost.

Evaluation

Accuracy of resolving 500 dialect words from dictionary, both categories:

# training data	Dialect Edit Distance	Edit Distance
500	39.6%	26.4%
1000	43.2%	28.2%
2000	46.0%	27.4%
5000	41.2%	27.2%

Accuracy of resolving 500 dialect words from dictionary, only phonologically related:

# training data	Dialect Edit Distance	Edit Distance
500	62.6%	40.8%
1000	63.6%	41.2%
2000	61.8%	37.8%
5000	64.8%	37.0%

Outline

Used Methodology

One possible way: extracting meaningful terms separately as tags, linking documents with same tags
Problem: synonymy.

Our approach: create hierarchy of related documents, attach word tags that are meaningful later for browsing
Used algorithm: Hierarchical Co-clustering (Xu and Ma)

Co-clustering

Document-clustering: based on word-similarity
Word-clustering: based on document-similarity

Co-clustering: iteratively

- cluster documents based on clustered words, and
- recluster words based on clustered documents

Result: Clusters of related documents, together with words relevant to these documents

Hierarchical Co-clustering

Hierarchical Co-clustering by Xu and Ma:
Spectrum-based, top down (divisive) Co-clustering
Determination # clusters by eigengap

Evaluation

Eigengap-method resulted in wrong number of clusters: very low recall.
Causes:

- Small corpus
- Low overlap in texts, clustering created by random word identities

Solutions

- Use different clustering method
- Create bigger corpus, once project runs
- Use document expansion to alleviate dialect word problem

The End

Questions?