Semantic Models for Style-based Text Clustering

A. Leoncini*, F. Sangiacomo*, C. Peretti*, S. Argentesi*,
E. Cambria◇, and R. Zunino*

*Dept. of Biophysical and Electronic Engineering
University of Genoa, Italy

◇Temasek Laboratories
National University of Singapore, Singapore

IEEE International Conference on Semantic Computing
September 19-21, 2011
Stanford University, Palo Alto, CA, USA

Outline

Purpose of this work:

- Tuning the use of semantic networks and defining a novel semantic-based metrics to increase accuracy in text clustering.

The outline of this work is:

- Introduction;
- Clustering Engine and Semantic Network;
- Semantic Style-based Document representation;
- Results and conclusions.

Outline

Purpose of this work:

- Tuning the use of semantic networks and defining a novel semantic-based metrics to increase accuracy in text clustering.

The outline of this work is:

- Introduction;
- Clustering Engine and Semantic Network;
- Semantic Style-based Document representation;
- Results and conclusions.

Dealing with massive corpora of documents

- A critical task in knowledge acquisition and intelligence gathering is **to organize** in a structured way a large amount of unstructured data (e.g. text documents).
- Structured data helps considerably analysts to **quickly collect dataset content**.
- When tagging facility is not available, unsupervised clustering algorithm can be an effective approach.
- Possible application: automated grouping of business emails in order to find malicious ones.

Dealing with massive corpora of documents

- A critical task in knowledge acquisition and intelligence gathering is **to organize** in a structured way a large amount of unstructured data (e.g. text documents).
- Structured data helps considerably analysts to **quickly collect dataset content**.
- When tagging facility is not available, unsupervised clustering algorithm can be an effective approach.
- Possible application: automated grouping of business emails in order to find malicious ones.

Dealing with massive corpora of documents

- A critical task in knowledge acquisition and intelligence gathering is **to organize** in a structured way a large amount of unstructured data (e.g. text documents).
- Structured data helps considerably analysts to **quickly collect dataset content**.
- When tagging facility is not available, unsupervised clustering algorithm can be an effective approach.
- Possible application: automated grouping of business emails in order to find malicious ones.

Dealing with massive corpora of documents

- A critical task in knowledge acquisition and intelligence gathering is **to organize** in a structured way a large amount of unstructured data (e.g. text documents).
- Structured data helps considerably analysts to **quickly collect dataset content**.

- When tagging facility is not available, unsupervised clustering algorithm can be an effective approach.
- Possible application: automated grouping of business emails in order to find malicious ones.

Semantic networks

Several works about document clustering proved that the use of semantic networks can overcome the bare word analysis, in terms of categorization accuracy.

- A common document representation, the Vector Space Model, considers different words as different vector dimensions.
- Semantic links like synonymy, hypernymy help to collapse different words into the same concept, or making links between different concepts.
- Document representation can embed information about semantic links; this allows the clustering process to consider also concepts other than document words.

Semantic networks

Several works about document clustering proved that the use of semantic networks can overcome the bare word analysis, in terms of categorization accuracy.

- A common document representation, the Vector Space Model, considers different words as different vector dimensions.
- Semantic links like synonymy, hypernymy help to collapse different words into the same concept, or making links between different concepts.
- Document representation can embed information about semantic links; this allows the clustering process to consider also concepts other than document words.

Semantic networks

Several works about document clustering proved that the use of semantic networks can overcome the bare word analysis, in terms of categorization accuracy.

- A common document representation, the Vector Space Model, considers different words as different vector dimensions.
- Semantic links like synonymy, hypernymy help to collapse different words into the same concept, or making links between different concepts.
- Document representation can embed information about semantic links; this allows the clustering process to consider also concepts other than document words.

The Document Clustering Framework

Preliminary actions:

- The reference document clustering engine, SLAIR, is a versatile framework based on Kernel K-Means.
- EuroWordNet lexical database was plugged into SLAIR, to enable semantic capabilities.

The clustering process is based on the following phases:

- Stop words removal and stemming (every word replaced with its base form);
- Semantic descriptor computation, querying EuroWordNet;
- Hierarchical Kernel K-means execution, with dynamic branch creation in the clusters tree.

The Document Clustering Framework

Preliminary actions:

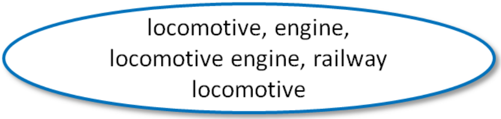
- The reference document clustering engine, SLAIR, is a versatile framework based on Kernel K-Means.
- EuroWordNet lexical database was plugged into SLAIR, to enable semantic capabilities.

The clustering process is based on the following phases:

- Stop words removal and stemming (every word replaced with its base form);
- Semantic descriptor computation, querying EuroWordNet;
- Hierarchical Kernel K-means execution, with dynamic branch creation in the clusters tree.

EuroWordNet semantic network

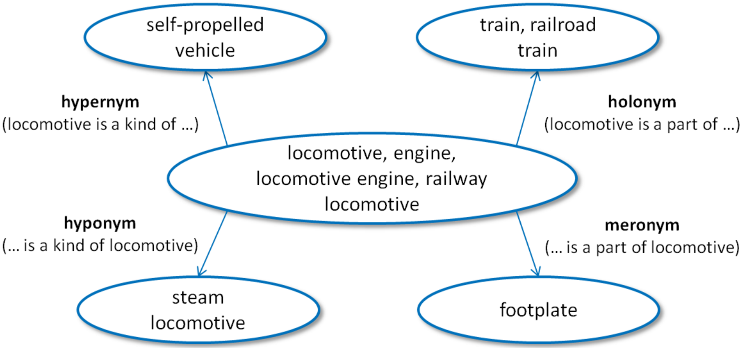
- EuroWordNet is an ontology containing words and semantic links.
- The set of words linked by a **synonym** relation is called **synset**, as shown in the following figure.



- Synsets are linked to other synsets with relations like **antonym** (opposite concept), **hypernym** (more general concept), and others.

EuroWordNet semantic network

Following figure shows the basic structure of EuroWordNet.



Semantic network model

- The semantic network model includes sets of terms that represents concepts.
- Mappings between terms and concepts are **many–to–many**.
- Lexical matrix (with $E_{i,j} \in \{0,1\}$):

$$
\begin{array}{cc|ccccc|}
 & & \multicolumn{5}{c}{\text{Terms}} \\
 & & t_1 & t_2 & t_3 & \dots & t_n \\
\hline
 & C_1 & E_{1,1} & E_{1,2} & & & \\
 & C_2 & & E_{2,2} & & & \\
\text{Concepts} & C_3 & & & E_{3,3} & & \\
 & \vdots & & & & \ddots & \\
 & C_m & & & & & E_{n,m} \\
\hline
\end{array}
$$

- These mappings possibly cause emersion of irrelevant concepts. Word Sense Disambiguation can help here to select the appropriate concept.

Semantic network model

- The semantic network model includes sets of terms that represents concepts.
- Mappings between terms and concepts are **many–to–many**.
- Lexical matrix (with $E_{i,j} \in \{0,1\}$):

$$
\begin{array}{ll|ccccc|}
 & & \multicolumn{5}{c}{\text{Terms}} \\
 & & t_1 & t_2 & t_3 & \dots & t_n \\
\hline
 & C_1 & E_{1,1} & E_{1,2} & & & \\
 & C_2 & & E_{2,2} & & & \\
\text{Concepts} & C_3 & & & E_{3,3} & & \\
 & \vdots & & & & \ddots & \\
 & C_m & & & & & E_{n,m} \\
\hline
\end{array}
$$

- These mappings possibly cause emersion of irrelevant concepts. Word Sense Disambiguation can help here to select the appropriate concept.

Semantic network model

- The semantic network model includes sets of terms that represents concepts.
- Mappings between terms and concepts are **many–to–many**.
- Lexical matrix (with $E_{i,j} \in \{0,1\}$):

$$
\begin{array}{ccc}
 & & \text{Terms} \\
 & & \begin{array}{ccccc} t_1 & t_2 & t_3 & \dots & t_n \end{array} \\
\text{Concepts} & \begin{array}{c} C_1 \\ C_2 \\ C_3 \\ \vdots \\ C_m \end{array} & \left|\begin{array}{ccccc} E_{1,1} & E_{1,2} & & & \\ & E_{2,2} & & & \\ & & E_{3,3} & & \\ & & & \ddots & \\ & & & & E_{n,m} \end{array}\right|
\end{array}
$$

- These mappings possibly cause emersion of irrelevant concepts. Word Sense Disambiguation can help here to select the appropriate concept.

Semantic network model

Introducing the basic Vector Space Model:

- A dictionary with terms from all the documents $T = \{t_j; j = 1, ..., n_T\}$
- A document D is expressed as a vector of term weights $\mathbf{v} = \{w_j; j = 1, ..., n_T\}$

A semantic network is assumed to support two operations:

- **Remapping**: semantic links remap $\mathbf{v}$ into a new vector $\mathbf{z}$ that spans the C concepts space rather than T space.
- **Compression**: shrinking vector $\mathbf{z}$ is allowed from hierarchic links between concepts.

Semantic network model

Introducing the basic Vector Space Model:

- A dictionary with terms from all the documents
  $T = \{t_j; j = 1, ..., n_T\}$
- A document D is expressed as a vector of term weights
  $\mathbf{v} = \{w_j; j = 1, ..., n_T\}$

A semantic network is assumed to support two operations:

- **Remapping**: semantic links remap $\mathbf{v}$ into a new vector $\mathbf{z}$ that spans the C concepts space rather than T space.
- **Compression**: shrinking vector $\mathbf{z}$ is allowed from hierarchic links between concepts.

Semantic-based document representation

Given desired number of meanings γ, remapping from T to C is provided by:

$$\mathrm{syn}(t_j, \gamma) = \mathbf{S}^{(j)}$$

- $\mathbf{S}^{(j)}$ is the concepts set of j-th term;
- $\dim(\mathbf{S}^{(j)}) \leq \gamma$ depending of the number of synonyms of j-th term stored in the semantic network.

The operator used to extract a set of hypernyms $\mathbf{H}^{(j)}$, given ξ as desired number of hierarchy steps, is:

$$\mathrm{hyp}(\mathbf{S}^{(j)}, \xi) = \mathbf{H}^{(j)}$$

- $\mathbf{H}^{(j)}$ is the hypernyms set of j-th term;
- $\dim(\mathbf{H}^{(j)}) \leq \xi$ depending on the number of hypernyms for j-th term stored in the semantic network.

Semantic-based document representation

Given desired number of meanings γ, remapping from T to C is provided by:

$$\mathrm{syn}(t_j, \gamma) = \mathbf{S}^{(j)}$$

- $\mathbf{S}^{(j)}$ is the concepts set of j-th term;
- $\dim(\mathbf{S}^{(j)}) \leq \gamma$ depending of the number of synonyms of j-th term stored in the semantic network.

The operator used to extract a set of hypernyms $\mathbf{H}^{(j)}$, given ξ as desired number of hierarchy steps, is:

$$\mathrm{hyp}(\mathbf{S}^{(j)}, \xi) = \mathbf{H}^{(j)}$$

- $\mathbf{H}^{(j)}$ is the hypernyms set of j-th term;
- $\dim(\mathbf{H}^{(j)}) \leq \xi$ depending on the number of hypernyms for j-th term stored in the semantic network.

Semantic-based document representation

- Given γ and ξ, for every term t_j there's a set $\mathbf{C}^{(j)} = \mathbf{S}^{(j)} \cup \mathbf{H}^{(j)}$.
- Considering terms from all the documents, is obtained a set of concepts $\mathbf{C}^*$ with all concepts, without duplicates, ordered by occurrences.
- All $\mathbf{v}$ terms are replaced with first corresponding concept cointained in the $\mathbf{C}^*$ set, building the vector $\mathbf{z}$.
- Parameters γ and ξ are crucial to the overall model effectiveness:
  - γ specifies how many meanings (i.e. synonyms) are retrieved for a single term;
  - ξ sets a level of abstraction, that is, the largest number of hierarchy levels that may separate a concept from one of its hypernyms.

Semantic-based document representation

- Given γ and ξ, for every term t_j there's a set $\mathbf{C}^{(j)} = \mathbf{S}^{(j)} \cup \mathbf{H}^{(j)}$.
- Considering terms from all the documents, is obtained a set of concepts $\mathbf{C}^*$ with all concepts, without duplicates, ordered by occurrences.
- All $\mathbf{v}$ terms are replaced with first corresponding concept cointained in the $\mathbf{C}^*$ set, building the vector $\mathbf{z}$.
- Parameters γ and ξ are crucial to the overall model effectiveness:
    - γ specifies how many meanings (i.e. synonyms) are retrieved for a single term;
    - ξ sets a level of abstraction, that is, the largest number of hierarchy levels that may separate a concept from one of its hypernyms.

Style-based document representation

Simple definition of **style**: the position of a term in the document.

- Every document is divided into Q sections;
- vector $\mathbf{p}$ represents hystograms of Q columns, one for every term in $\mathbf{z}$.

Example with $Q = 5$:

Semantic Style-based representation

This research proposes a distance between documents that includes two terms:

- frequency-based term: $\Delta^{(f)}(D_u, D_v) = ||\mathbf{z}(D_u) - \mathbf{z}(D_v)||^2$
- style-based term: $\Delta^{(s)}(D_u, D_v) = ||\mathbf{p}(D_u) - \mathbf{p}(D_v)||^2$

The eventual distance value stems from the linear combination of the two terms, with a balancing factor $\alpha \in [0, 1]$:

$$\Delta(D_u, D_v) = \alpha\Delta^{(f)}(D_u, D_v) + (1 - \alpha)\Delta^{(s)}(D_u, D_v)$$

Semantic Style-based representation

This research proposes a distance between documents that includes two terms:

- frequency-based term: $\Delta^{(f)}(D_u, D_v) = ||\mathbf{z}(D_u) - \mathbf{z}(D_v)||^2$
- style-based term: $\Delta^{(s)}(D_u, D_v) = ||\mathbf{p}(D_u) - \mathbf{p}(D_v)||^2$

The eventual distance value stems from the linear combination of the two terms, with a balancing factor $\alpha \in [0, 1]$:

$$\Delta(D_u, D_v) = \alpha\Delta^{(f)}(D_u, D_v) + (1 - \alpha)\Delta^{(s)}(D_u, D_v)$$

Experiments preamble

Datasets presented:

- Webcrawling dataset, 896 New York Times articles concerning three main topics;
- Enron5k dataset, 5,000 emails of five different authors;
- RCV1 subset, 13,832 Reuters news articles of four subjects;
- RCV2 subset, 12,594 Reuters Italian news of four subjects.

Quality criterion:

- categorization accuracy, measured over existing tagged corpora;
- comparing with well-known Vector Space (frequential) model.

Parameters chosen:

- number Q of document sections, to build $\mathbf{p}$ vector, is set to 5;
- distance balancing factor α is set to 0.2.

Experiments preamble

Datasets presented:

- Webcrawling dataset, 896 New York Times articles concerning three main topics;
- Enron5k dataset, 5,000 emails of five different authors;
- RCV1 subset, 13,832 Reuters news articles of four subjects;
- RCV2 subset, 12,594 Reuters Italian news of four subjects.

Quality criterion:

- categorization accuracy, measured over existing tagged corpora;
- comparing with well-known Vector Space (frequential) model.

Parameters chosen:

- number Q of document sections, to build $\mathbf{p}$ vector, is set to 5;
- distance balancing factor α is set to 0.2.

Experiments preamble

Datasets presented:

- Webcrawling dataset, 896 New York Times articles concerning three main topics;
- Enron5k dataset, 5,000 emails of five different authors;
- RCV1 subset, 13,832 Reuters news articles of four subjects;
- RCV2 subset, 12,594 Reuters Italian news of four subjects.

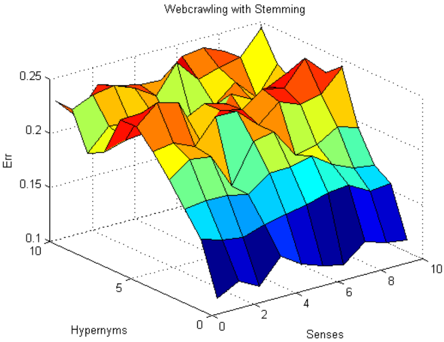
Quality criterion:

- categorization accuracy, measured over existing tagged corpora;
- comparing with well-known Vector Space (frequential) model.

Parameters chosen:

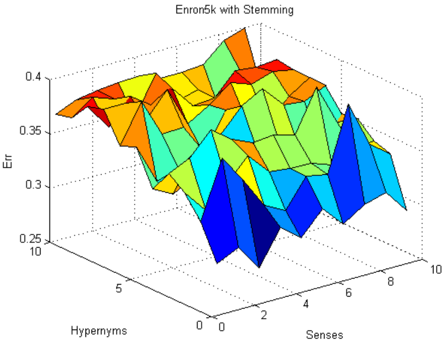
- number Q of document sections, to build $\mathbf{p}$ vector, is set to 5;
- distance balancing factor α is set to 0.2.

Experimental results



Webcrawling test: categorization error versus number of senses and hypernyms (differences between stemming phase disabled and enabled).

Experimental results



Enron5k test and RCV1 test: categorization error versus number of senses and hypernyms.

Experimental results

Average hypernyms chain and convenience of the stemming phase

Results	Datasets		
	Webcrawling	Enron5k	RCV1 subset
Vocabulary size	38260	35773	85782
Average hypernyms	5.18	5.11	5.02

Results	Datasets			
	Webcrawling	Enron5k	RCV1 subset	RCV2 subset
EWN hits	63.40 %	59.44 %	60.7 %	41.93 %
EWN hits with stemmer	80.67 %	69.01 %	75.04 %	57.13 %
Stemmer gain	17.27 %	9.57 %	14.34 %	15.20 %

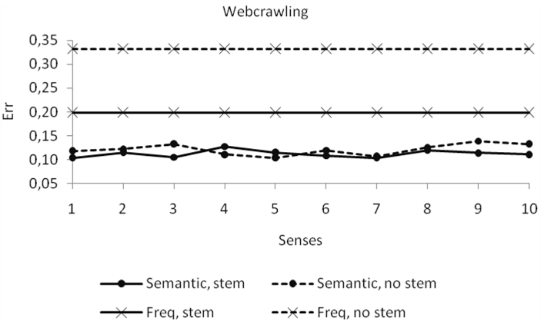
Experimental results

Average hypernyms chain and convenience of the stemming phase

Results	Datasets		
	Webcrawling	**Enron5k**	**RCV1 subset**
Vocabulary size	38260	35773	85782
Average hypernyms	5.18	5.11	5.02

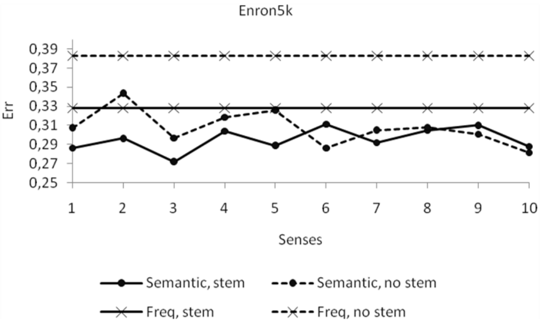
Results	Datasets			
	Webcrawling	**Enron5k**	**RCV1 subset**	**RCV2 subset**
EWN hits	63.40 %	59.44 %	60.7 %	41.93 %
EWN hits with stemmer	80.67 %	69.01 %	75.04 %	57.13 %
Stemmer gain	17.27 %	9.57 %	14.34 %	15.20 %

Experimental results



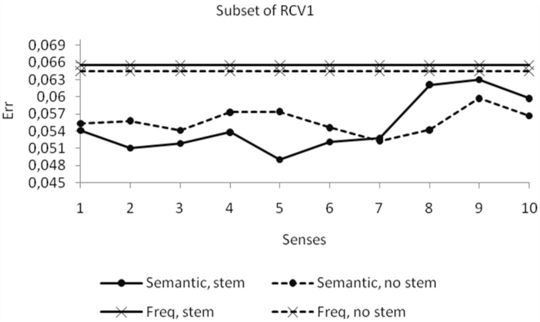
Categorization error obtained with the semantic model and with common frequential model, over Webcrawling dataset.

Experimental results



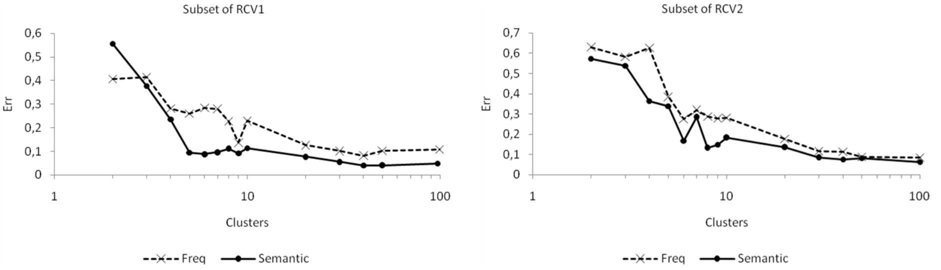
Categorization error using semantic model and frequential model, over the subset of Enron dataset.

Experimental results



Categorization error obtained using semantic and frequential models, over the subset of the RCV1 dataset.

Experimental results



Flat clustering, with K number of clusters in the range 2–100. Semantic model vs. Freq model, over the subset of RCV1 dataset and subset or RCV2 dataset (italian).

Conclusions and Future works

- Crucial novelty aspect of this work is the integration of a semantic-based representation and a hybrid frequency-stylistic metric into the kernel-based clustering engine.
- The main advantage is that dimensionality reduction is supported by external hidden information, i.e. the semantic knowledge.
- The style-based schema **always outperformed** the conventional approach. Best results were obtained by setting $\gamma = \mathbf{3}$ and $\xi = \mathbf{0}$, in which case the stemmer contributed profitably to the document categorization.
- Future works will study **Word Sense Disambiguation** to help choosing appropriate concept from terms, and language independent semantic networks.

Conclusions and Future works

- Crucial novelty aspect of this work is the integration of a semantic-based representation and a hybrid frequency-stylistic metric into the kernel-based clustering engine.
- The main advantage is that dimensionality reduction is supported by external hidden information, i.e. the semantic knowledge.
- The style-based schema **always outperformed** the conventional approach. Best results were obtained by setting $\gamma = \mathbf{3}$ and $\xi = \mathbf{0}$, in which case the stemmer contributed profitably to the document categorization.
- Future works will study **Word Sense Disambiguation** to help choosing appropriate concept from terms, and language independent semantic networks.

Conclusions and Future works

- Crucial novelty aspect of this work is the integration of a semantic-based representation and a hybrid frequency-stylistic metric into the kernel-based clustering engine.
- The main advantage is that dimensionality reduction is supported by external hidden information, i.e. the semantic knowledge.
- The style-based schema **always outperformed** the conventional approach. Best results were obtained by setting $\gamma = \mathbf{3}$ and $\xi = \mathbf{0}$, in which case the stemmer contributed profitably to the document categorization.
- Future works will study **Word Sense Disambiguation** to help choosing appropriate concept from terms, and language independent semantic networks.

Conclusions and Future works

- Crucial novelty aspect of this work is the integration of a semantic-based representation and a hybrid frequency-stylistic metric into the kernel-based clustering engine.
- The main advantage is that dimensionality reduction is supported by external hidden information, i.e. the semantic knowledge.
- The style-based schema **always outperformed** the conventional approach. Best results were obtained by setting $\gamma = \mathbf{3}$ and $\xi = \mathbf{0}$, in which case the stemmer contributed profitably to the document categorization.
- Future works will study **Word Sense Disambiguation** to help choosing appropriate concept from terms, and language independent semantic networks.