The Impact of Conceptualization on Text Classification

Shereen Albitar, Sébastien Fournier, and Bernard Espinasse

Université d'Aix marseille, LSIS, av. escadrille Normandie Niemen, 13397, Marseille, France
{shereen.albitar,sebastien.fournier,bernard.espinasse}@lsis.org

Abstract. Aiming at more efficient search on the Internet, it seems adequate to deploy classification techniques using semantic resources restricting this search to the user's domain of interest. In this work, we try to assess the impact of integrating semantic knowledge on text classification. This integration can be realized in different ways. The one we choose in this paper is the conceptualization. We examine the impact of the different conceptualization strategies on text classification using three traditional text classification methods: Rocchio, Support Vector Machines (SVMs) and Naïve Bayes (NB). We restrain our experimentation to the biomedical domain so conceptualization is applied on OHSUMED corpus, mapping terms in text to their corresponding concepts in UMLS Metathesaurus in order to take their meaning into consideration during text classification. Rocchio, SVMs, and NB are tested using different conceptualization strategies in order to evaluate their effect on classification. Preliminary results demonstrate promising improvements.

Keywords: Text Classification, Semantic classification, Information retrieval, Rocchio, SVMs, NB, Similarity measures, conceptualization.

1 Introduction

Nowadays and due to the explosive increase in published information on Internet, existing search engines seem to be unable to respond efficiently to user requests. This is often related to the traditional keyword-based indexing techniques neglecting search domain [1]. Aiming at more efficient and less time expensive search, it seems adequate to involve classification techniques in order to analyze the contents of search engines' answers, applying thorough filtering and ranking. Concerning information published on internet in Web pages, preprocessing treatments and content extraction are necessary to prepare this information for classification [2].

Text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc. Generally, text classification methods use syntactical and statistical models for text document representation. This applies to the most popular text classification methods such as: Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and so forth. Making a comparative study on the different traditional classification methods is out of the scope of this paper, for detailed comparisons please refer to [3, 4].

X.S. Wang et al. (Eds.): WISE 2012, LNCS 7651, pp. 326–339, 2012.

These models suffer the lack of sense in resulting representations ignoring all semantics that reside in the original text that can help in text classification. However Vector-based (binary or TF/IDF) representations used by preceding methods permit semantic integration or "Conceptualization" that enriches document representation model using background knowledge bases [5, 6].

In this work, according to experiments, we try to estimate the impact of different text conceptualization strategies on traditional methods. These experiments are realized particularly in the biomedical domain on the OHSUMED corpus, using domain specific knowledge base UMLS. Many works have already tried to improve biomedical text representation for better classification using UMLS [7, 8] or MESH [9, 10]. To the best of our knowledge, few works investigated in details the gain of integrating these semantic resources in biomedical text classification process and more particularly, the impact of conceptualization.

Next section presents how semantic resources (thesaurus or ontologies) can be taken into account during text classification, through text conceptualization task, leading to a “semantic” text classification. In third section, we apply Rocchio (with different similarity measures) SVMs and NB methods to the original and the conceptualized OHSUMED corpus. This conceptualization is realized using UMLS (Unified Medical Language System) Metathesaurus and MetaMap tool. The forth section presents some preliminary results using these methods. Conceptualization effects on the performance of the studied methods, according to different conceptualization strategies are discussed. Then we expose some related works. Finally, we conclude with an assessment of our work, followed by different research perspectives.

2 Text Conceptualization Task

In order to overcome the previous limitations of traditional text representation models, semantic resources such as thesaurus or ontologies, can be used to replace term-based representation by a concept-based one. Thus, text classification using conceptualized vectors is so called "Semantic Classification". This section presents text conceptualization task, introducing different possible conceptualization and disambiguation strategies.

Conceptualization is the process of moving from terms literally occurring in treated text to their semantically corresponding concepts or senses in semantic resources that might permit better classification results. As an example of semantic resources that might be used for conceptualization: Wordnet, Wikipedia and other domain specific resources usually called domain ontologies such as UMLS thesaurus in the medical domain. In general, text conceptualization is realized in two steps:

- Analyze text in order to find candidate terms for term to concept mapping.
- Search for corresponding concepts related to candidate terms, and then the integration of these concepts in text producing the final conceptualized text.

2.1 Text Conceptualization Strategies

Three different strategies can be used for text conceptualization:

- *Adding Concepts*: in this strategy the original text is extended and corresponding concepts are added.
- *Partial Conceptualization*: in this strategy terms are substituted by corresponding concepts. Terms having no related concepts are reserved in the text.
- *Complete Conceptualization*: similarly to Partial Conceptualization, in this strategy terms are substituted by concepts whereas remaining terms are eliminated from the final text.

The second strategy seems to be the most appropriate one as it removes no term before replacing it with a related concept so no original feature is removed from the text (compared to the third one), and no extra feature is added (compared to the first one) resulting in minimized efficiency effects. Yet, the classification method has to be adapted to hybrid (concepts + terms) representation.

2.2 Disambiguation Strategies

While searching polysemic term corresponding concepts in semantic resources, multiple matches are detected and introducing some ambiguities in final document representation. For example: the term "Book" signifies in English a book and also a reservation (Ticket, accommodations, etc.). Three strategies for disambiguation can be used:

- *All*: this strategy accepts all candidate concepts as matches for the considered term.
- *First*: this strategy accepts the most frequently used concept among candidates using language statistics.
- *Context*: this strategy accepts the candidate concepts having the most similar semantic context compared to the term's context in the document.

The first strategy, despite being the simplest, is the least reliable as it accepts all candidate concepts without choosing a specific sense of the term. In cases where a term is used in the document signifying its rarely used sense, the second strategy gives bad decision. Despite its complexity, the last strategy seems to be more accurate as concept context can be derived from semantic resources using: concepts definition, its descriptive terms or from text corpus.

3 Platform for Conceptualized Text Classification

This section presents an experimental platform for assessing the impact of different conceptualization strategies on text classification, using three traditional text classification methods: Rocchio, SVMs and NB. This platform is illustrated in figure 1. First we present the components constituting this platform: the biomedical knowledge base

UMLS (Unified Medical Language System), and the MetaMap tool that deploys UMLS in order to realize text to concept mapping or conceptualization on the OHSUMED corpus. Conceptualized text is then transformed into vectors of terms and/or concepts depending on the used conceptualization strategy.

During Training phase the training corpus is prepared for training the classifier resulting in a classification model whereas during the Classification phase the test corpus is prepared in order to attribute classes to its document by the classifier using the learned classification model.

This architecture is modular and generic so different components can be modified and even replaced. In this work we use three different traditional classification algorithms, Rocchio, NB and SVMs, to realize Training and Classification phases.

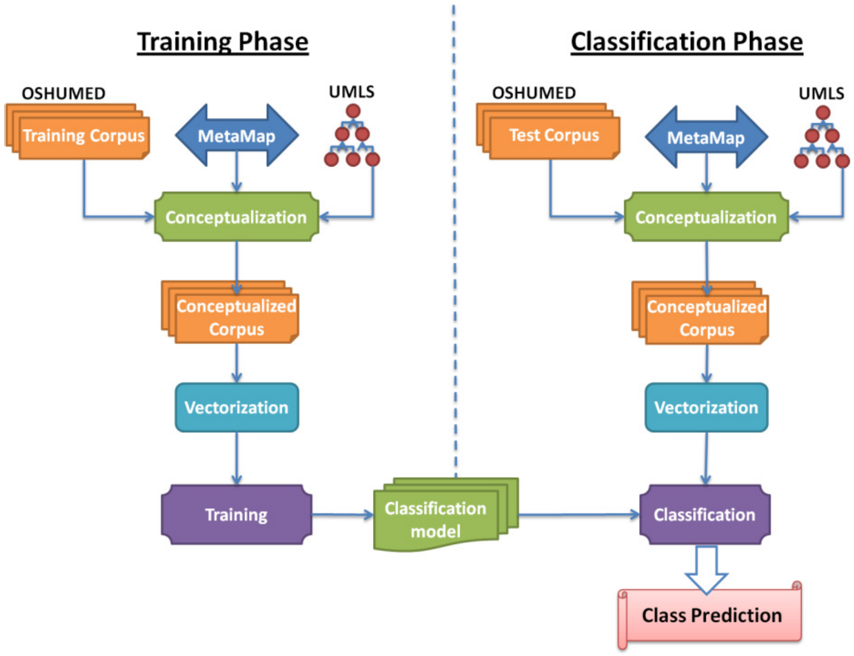


Fig. 1. The architecture of a platform for conceptualized text classification using MetaMap and UMLS

3.1 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [11] was developed in order to model the language of biomedicine and health. UMLS' knowledge sources enhance the development of information systems in the biomedical domain.

The UMLS knowledge base consists of three main resources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus is a multilingual vocabulary database of biomedical concepts, their names, their attributes and the relations among them. This database organizes concepts of the various source vocabularies (like Mesh, SNOMED-CT, etc.) according to their senses grouping common concepts together. Concepts and relations among them are assigned at least one type from the Semantic Network. Indeed, the Semantic Network provides a higher level of abstraction through concept and relation categorization in inter-related Types constituting a network of 133 semantic types and 54 relationships. The SPECIALIST lexicon contains a large variety of general as well as medical terms and words.

3.2 MetaMap Tool

In addition to the UMLS knowledge resources, many tools are developed and provided by NLM in order to facilitate deploying these sources for medical information system developers. In this work we are interested particularly in MetaMap [12]. The major goal of MetaMap developers was to improve biomedical text retrieval using UMLS Metathesaurus. Indeed, MetaMap can discover the links between biomedical text and the knowledge in the Metathesaurus.

This mapping is the result of a rigorous linguistic analysis of each phrase of the text: First the text is tokenized and phrase boundaries are identified, then part of-speech-tags are added. Second, the Specialist lexicon and the shallow parser are used to analyze syntactically these phrases. Finally, different candidates are identified in the Metathesaurus and then final mappings combining these candidates are evaluated resulting in confidence scores for each mapping. In cases were ambiguities are detected, MetaMap keeps the most semantically similar mappings to the surrounding text following the context strategy (section 3.1). Text conceptualization using UMLS as a semantic resource, if compared to other more generic semantic resources such as WordNet, is more relevant in the medical domain. This allows the mapping between text and related specific concepts in UMLS.

3.3 OHSUMED Corpus

OHSUMED corpus [13] is composed of abstracts of biomedical articles of the year 1991 retrieved from the MEDLINE database indexed using MeSh (Medical Subject Headings) [14]. The first 20000 documents of this database were selected and categorized using 23 sub-concepts of the Mesh concept "Disease".

The corpus is divided in Training and Test sets, so experimentations are realized in two phases: Training and Test. In this work, we restricted this corpus to the five most frequent classes [15]. Training is realized on the corpus and so five class centroïds are calculated for each of the classes listed in Table 1.

Table 1. OHSUMED Corpus

Category	Description
C04	Neoplasms
C23	Pathological Conditions, Signs and
C06	Digestive System Diseases
C14	Cardiovascular Diseases
C20	Immune System Diseases

4 Experiments and Results

Using the previously presented platform, we have performed experiments in order to evaluate the effect of conceptualization on the classification process using three traditional classification methods: SVMs, Rocchio and NB.

In this section we present first the text conceptualization task performed on the OHSUMED corpus according to UMLS Metathesaurus, according to different strategies, and using the MetaMap tool. Then we present classification results obtained with each of the tree traditional classification methods (Rocchio, SVMs and NB) for each of these conceptualization strategies. Finally we analyze and discuss these results obtained.

4.1 Text Conceptualization Task

During the conceptualization task, different strategies can be implemented as previously described (adding concepts, partial conceptualization and complete conceptualization). Futhermore, according to MetaMap text-to-concept matching results, we can choose two complementary strategies:

- *Best concept strategy*. The best concept among several candidate concepts matched to the text. This depends on a matching score computed by MetaMap [12].
- *All concepts strategy*. All candidate concepts are kept.

Candidates resulting from matching have many properties. In this work we choose to use the concept name or the concept ID. In fact, during the tokenization step, the concept ID is considered as a single token so it stays intact. Concept names, being sometimes compound words, are cut during tokenization when applied on a text conceptualized using concept name strategy.

In this work, conceptualization is done using all combinations of the different strategies (12 combinations).

4.2 Classification Task

The use of MetaMap enables mapping text to UMLS concepts. Conceptualization is realized as a first step before vectorization. In our experiments, we tested all methods on the OHSUMED corpus before and after conceptualization that is realized according to twelve different conceptualization strategies.

Three traditional classification methods we used in experiments:

- Rocchio, using three different similarity measures: Cosine, Jaccard, KullbackLeibler [16].
- SVMs, using the library LIBSVM [17].
- NB, using the platform Weka [18].

In all experiments, methods are evaluated through Holdout Validation. F1-Measure [19] is considered as the criterion for performance comparison. This measure is given though the following equations:

$$\mathrm{Precision} = \frac{\mathrm{t_p}}{\mathrm{t_p} + \mathrm{f_p}} \tag{9}$$

$$\mathrm{Recall} = \frac{\mathrm{t_p}}{\mathrm{t_p} + \mathrm{f_n}} \tag{10}$$

$$\mathrm{F1 - Measure} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{11}$$

Considering a specific category *C*:

$\mathrm{t_p}$: The number of correctly classified documents in *C*.

$\mathrm{t_n}$: The number of correctly classified documents in other categories.

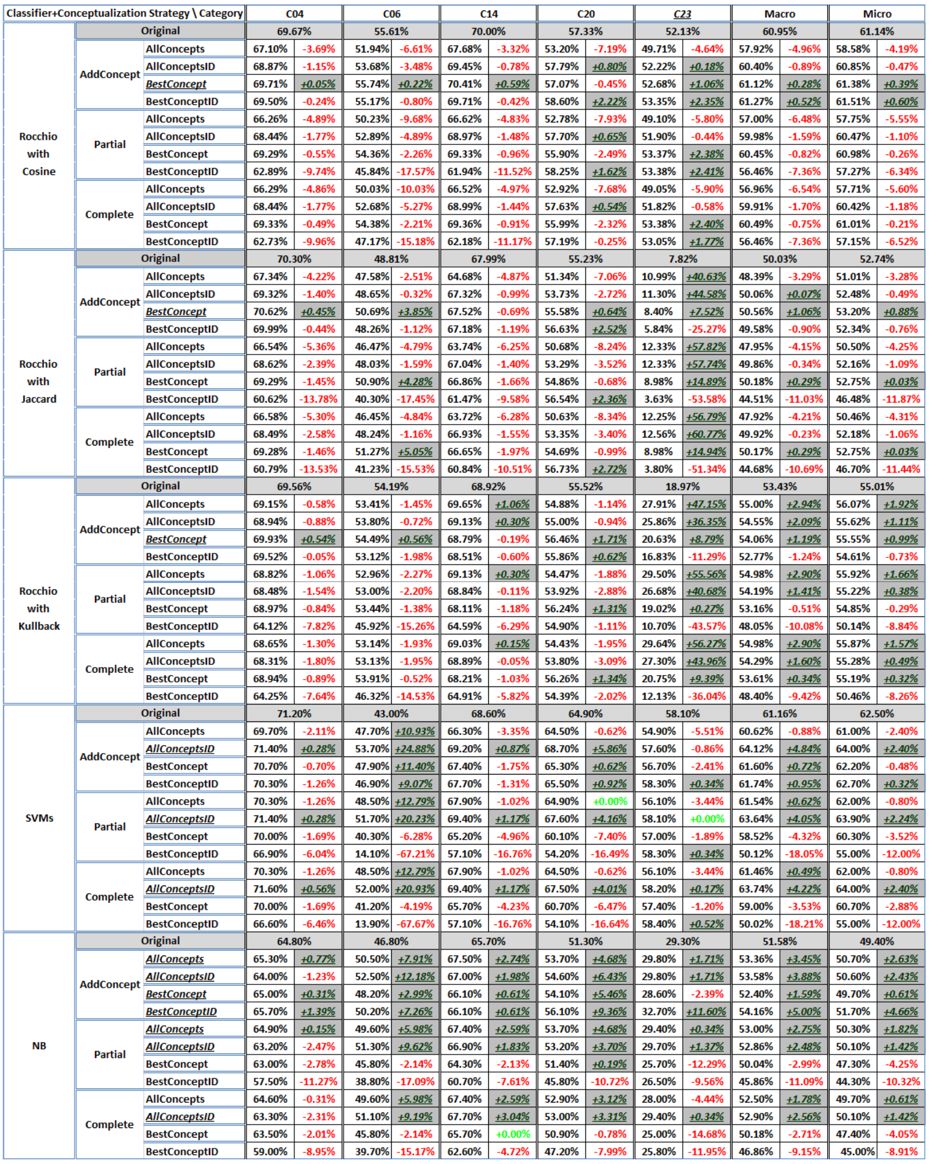
$\mathrm{f_p}$: The number of documents classified in *C* while they belong to other categories.

$\mathrm{f_n}$: The number of documents classified in other categories while they belong to *C*

4.3 Results Analysis

Table 2 shows the F1-measure obtained applying Rocchio (with three different similarity measures: Cosine, Jaccard and KullBack), SVMs and NB to conceptualized OHSUMED corpus, for each of five categories {C04, C06, C14, C20, C23}. First we test the five classification methods on *original* OHSUMED corpus without conceptualization. Then methods are tested using each of the twelve different conceptualization strategies that are introduced previously. The two last columns present Micro and Macro averaged F-measure obtained for each pair of classification methods and conceptualization strategies. In micro-averaging, F-measure is computed globally over all category documents, whereas in macro-averaging it is equal to the average of locally calculated F-measure for each class.

Table 2. Results of applying Rocchio to conceptualized OHSUMED



Classifier+Conceptualization Strategy \ Category			C04		C06		C14		C20		*C23*		Macro		Micro	
Rocchio with Cosine	Original		69.67%		55.61%		70.00%		57.33%		52.13%		60.95%		61.14%	
	AddConcept	AllConcepts	67.10%	-3.69%	51.94%	-6.61%	67.68%	-3.32%	53.20%	-7.19%	49.71%	-4.64%	57.92%	-4.96%	58.58%	-4.19%
		AllConceptsID	68.87%	-1.15%	53.68%	-3.48%	69.45%	-0.78%	57.79%	*+0.80%*	52.22%	*+0.18%*	60.40%	-0.89%	60.85%	-0.47%
		BestConcept	69.71%	*+0.05%*	55.74%	*+0.22%*	70.41%	*+0.59%*	57.07%	-0.45%	52.68%	*+1.06%*	61.12%	*+0.28%*	61.38%	*+0.39%*
		BestConceptID	69.50%	-0.24%	55.17%	-0.80%	69.71%	-0.42%	58.60%	*+2.22%*	53.35%	*+2.35%*	61.27%	*+0.52%*	61.51%	*+0.60%*
	Partial	AllConcepts	66.26%	-4.89%	50.23%	-9.68%	66.62%	-4.83%	52.78%	-7.93%	49.10%	-5.80%	57.00%	-6.48%	57.75%	-5.55%
		AllConceptsID	68.44%	-1.77%	52.89%	-4.89%	68.97%	-1.48%	57.70%	*+0.65%*	51.90%	-0.44%	59.98%	-1.59%	60.47%	-1.10%
		BestConcept	69.29%	-0.55%	54.36%	-2.26%	69.33%	-0.96%	55.90%	-2.49%	53.37%	*+2.38%*	60.45%	-0.82%	60.98%	-0.26%
		BestConceptID	62.89%	-9.74%	45.84%	-17.57%	61.94%	-11.52%	58.25%	*+1.62%*	53.38%	*+2.41%*	56.46%	-7.36%	57.27%	-6.34%
	Complete	AllConcepts	66.29%	-4.86%	50.03%	-10.03%	66.52%	-4.97%	52.92%	-7.68%	49.05%	-5.90%	56.96%	-6.54%	57.71%	-5.60%
		AllConceptsID	68.44%	-1.77%	52.68%	-5.27%	68.99%	-1.44%	57.63%	*+0.54%*	51.82%	-0.58%	59.91%	-1.70%	60.42%	-1.18%
		BestConcept	69.33%	-0.49%	54.38%	-2.21%	69.36%	-0.91%	55.99%	-2.32%	53.38%	*+2.40%*	60.49%	-0.75%	61.01%	-0.21%
		BestConceptID	62.73%	-9.96%	47.17%	-15.18%	62.18%	-11.17%	57.19%	-0.25%	53.05%	*+1.77%*	56.46%	-7.36%	57.15%	-6.52%
Rocchio with Jaccard	Original		70.30%		48.81%		67.99%		55.23%		7.82%		50.03%		52.74%	
	AddConcept	AllConcepts	67.34%	-4.22%	47.58%	-2.51%	64.68%	-4.87%	51.34%	-7.06%	10.99%	*+40.63%*	48.39%	-3.29%	51.01%	-3.28%
		AllConceptsID	69.32%	-1.40%	48.65%	-0.32%	67.32%	-0.99%	53.73%	-2.72%	11.30%	*+44.58%*	50.06%	*+0.07%*	52.48%	-0.49%
		BestConcept	70.62%	*+0.45%*	50.69%	*+3.85%*	67.52%	-0.69%	55.58%	*+0.64%*	8.40%	*+7.52%*	50.56%	*+1.06%*	53.20%	*+0.88%*
		BestConceptID	69.99%	-0.44%	48.26%	-1.12%	67.18%	-1.19%	56.63%	*+2.52%*	5.84%	-25.27%	49.58%	-0.90%	52.34%	-0.76%
	Partial	AllConcepts	66.54%	-5.36%	46.47%	-4.79%	63.74%	-6.25%	50.68%	-8.24%	12.33%	*+57.82%*	47.95%	-4.15%	50.50%	-4.25%
		AllConceptsID	68.62%	-2.39%	48.03%	-1.59%	67.04%	-1.40%	53.29%	-3.52%	12.33%	*+57.74%*	49.86%	-0.34%	52.16%	-1.09%
		BestConcept	69.29%	-1.45%	50.90%	*+4.28%*	66.86%	-1.66%	54.86%	-0.68%	8.98%	*+14.89%*	50.18%	*+0.29%*	52.75%	*+0.03%*
		BestConceptID	60.62%	-13.78%	40.30%	-17.45%	61.47%	-9.58%	56.54%	*+2.36%*	3.63%	-53.58%	44.51%	-11.03%	46.48%	-11.87%
	Complete	AllConcepts	66.58%	-5.30%	46.45%	-4.84%	63.72%	-6.28%	50.63%	-8.34%	12.25%	*+56.79%*	47.92%	-4.21%	50.46%	-4.31%
		AllConceptsID	68.49%	-2.58%	48.24%	-1.16%	66.93%	-1.55%	53.35%	-3.40%	12.56%	*+60.77%*	49.92%	-0.23%	52.18%	-1.06%
		BestConcept	69.28%	-1.46%	51.27%	*+5.05%*	66.65%	-1.97%	54.69%	-0.99%	8.98%	*+14.94%*	50.17%	*+0.29%*	52.75%	*+0.03%*
		BestConceptID	60.79%	-13.53%	41.23%	-15.53%	60.84%	-10.51%	56.73%	*+2.72%*	3.80%	-51.34%	44.68%	-10.69%	46.70%	-11.44%
Rocchio with Kullback	Original		69.56%		54.19%		68.92%		55.52%		18.97%		53.43%		55.01%	
	AddConcept	AllConcepts	69.15%	-0.58%	53.41%	-1.45%	69.65%	*+1.06%*	54.88%	-1.14%	27.91%	*+47.15%*	55.00%	*+2.94%*	56.07%	*+1.92%*
		AllConceptsID	68.94%	-0.88%	53.80%	-0.72%	69.13%	*+0.30%*	55.00%	-0.94%	25.86%	*+36.35%*	54.55%	*+2.09%*	55.62%	*+1.11%*
		BestConcept	69.93%	*+0.54%*	54.49%	*+0.56%*	68.79%	-0.19%	56.46%	*+1.71%*	20.63%	*+8.79%*	54.06%	*+1.19%*	55.55%	*+0.99%*
		BestConceptID	69.52%	-0.05%	53.12%	-1.98%	68.51%	-0.60%	55.86%	*+0.62%*	16.83%	-11.29%	52.77%	-1.24%	54.61%	-0.73%
	Partial	AllConcepts	68.82%	-1.06%	52.96%	-2.27%	69.13%	*+0.30%*	54.47%	-1.88%	29.50%	*+55.56%*	54.98%	*+2.90%*	55.92%	*+1.66%*
		AllConceptsID	68.48%	-1.54%	53.00%	-2.20%	68.84%	-0.11%	53.92%	-2.88%	26.68%	*+40.68%*	54.19%	*+1.41%*	55.22%	*+0.38%*
		BestConcept	68.97%	-0.84%	53.44%	-1.38%	68.11%	-1.18%	56.24%	*+1.31%*	19.02%	*+0.27%*	53.16%	-0.51%	54.85%	-0.29%
		BestConceptID	64.12%	-7.82%	45.92%	-15.26%	64.59%	-6.29%	54.90%	-1.11%	10.70%	-43.57%	48.05%	-10.08%	50.14%	-8.84%
	Complete	AllConcepts	68.65%	-1.30%	53.14%	-1.93%	69.03%	*+0.15%*	54.43%	-1.95%	29.64%	*+56.27%*	54.98%	*+2.90%*	55.87%	*+1.57%*
		AllConceptsID	68.31%	-1.80%	53.13%	-1.95%	68.89%	-0.05%	53.80%	-3.09%	27.30%	*+43.96%*	54.29%	*+1.60%*	55.28%	*+0.49%*
		BestConcept	68.94%	-0.89%	53.91%	-0.52%	68.21%	-1.03%	56.26%	*+1.34%*	20.75%	*+9.39%*	53.61%	*+0.34%*	55.19%	*+0.32%*
		BestConceptID	64.25%	-7.64%	46.32%	-14.53%	64.91%	-5.82%	54.39%	-2.02%	12.13%	-36.04%	48.40%	-9.42%	50.46%	-8.26%
SVMs	Original		71.20%		43.00%		68.60%		64.90%		58.10%		61.16%		62.50%	
	AddConcept	AllConcepts	69.70%	-2.11%	47.70%	*+10.93%*	66.30%	-3.35%	64.50%	-0.62%	54.90%	-5.51%	60.62%	-0.88%	61.00%	-2.40%
		AllConceptsID	71.40%	*+0.28%*	53.70%	*+24.88%*	69.20%	*+0.87%*	68.70%	*+5.86%*	57.60%	-0.86%	64.12%	*+4.84%*	64.00%	*+2.40%*
		BestConcept	70.70%	-0.70%	47.90%	*+11.40%*	67.40%	-1.75%	65.30%	*+0.62%*	56.70%	-2.41%	61.60%	*+0.72%*	62.20%	-0.48%
		BestConceptID	70.30%	-1.26%	46.90%	*+9.07%*	67.70%	-1.31%	65.50%	*+0.92%*	58.30%	*+0.34%*	61.74%	*+0.95%*	62.70%	*+0.32%*
	Partial	AllConcepts	70.30%	-1.26%	48.50%	*+12.79%*	67.90%	-1.02%	64.90%	+0.00%	56.10%	-3.44%	61.54%	*+0.62%*	62.00%	-0.80%
		AllConceptsID	71.40%	*+0.28%*	51.70%	*+20.23%*	69.40%	*+1.17%*	67.60%	*+4.16%*	58.10%	+0.00%	63.64%	*+4.05%*	63.90%	*+2.24%*
		BestConcept	70.00%	-1.69%	40.30%	-6.28%	65.20%	-4.96%	60.10%	-7.40%	57.00%	-1.89%	58.52%	-4.32%	60.30%	-3.52%
		BestConceptID	66.90%	-6.04%	14.10%	-67.21%	57.10%	-16.76%	54.20%	-16.49%	58.30%	*+0.34%*	50.12%	-18.05%	55.00%	-12.00%
	Complete	AllConcepts	70.30%	-1.26%	48.50%	*+12.79%*	67.90%	-1.02%	64.50%	-0.62%	56.10%	-3.44%	61.46%	*+0.49%*	62.00%	-0.80%
		AllConceptsID	71.60%	*+0.56%*	52.00%	*+20.93%*	69.40%	*+1.17%*	67.50%	*+4.01%*	58.20%	*+0.17%*	63.74%	*+4.22%*	64.00%	*+2.40%*
		BestConcept	70.00%	-1.69%	41.20%	-4.19%	65.70%	-4.23%	60.70%	-6.47%	57.40%	-1.20%	59.00%	-3.53%	60.70%	-2.88%
		BestConceptID	66.60%	-6.46%	13.90%	-67.67%	57.10%	-16.76%	54.10%	-16.64%	58.40%	*+0.52%*	50.02%	-18.21%	55.00%	-12.00%
NB	Original		64.80%		46.80%		65.70%		51.30%		29.30%		51.58%		49.40%	
	AddConcept	*AllConcepts*	65.30%	*+0.77%*	50.50%	*+7.91%*	67.50%	*+2.74%*	53.70%	*+4.68%*	29.80%	*+1.71%*	53.36%	*+3.45%*	50.70%	*+2.63%*
		AllConceptsID	64.00%	-1.23%	52.50%	*+12.18%*	67.00%	*+1.98%*	54.60%	*+6.43%*	29.80%	*+1.71%*	53.58%	*+3.88%*	50.60%	*+2.43%*
		BestConcept	65.00%	*+0.31%*	48.20%	*+2.99%*	66.10%	*+0.61%*	54.10%	*+5.46%*	28.60%	-2.39%	52.40%	*+1.59%*	49.70%	*+0.61%*
		BestConceptID	65.70%	*+1.39%*	50.20%	*+7.26%*	66.10%	*+0.61%*	56.10%	*+9.36%*	32.70%	*+11.60%*	54.16%	*+5.00%*	51.70%	*+4.66%*
	Partial	*AllConcepts*	64.90%	*+0.15%*	49.60%	*+5.98%*	67.40%	*+2.59%*	53.70%	*+4.68%*	29.40%	*+0.34%*	53.00%	*+2.75%*	50.30%	*+1.82%*
		AllConceptsID	63.20%	-2.47%	51.30%	*+9.62%*	66.90%	*+1.83%*	53.20%	*+3.70%*	29.70%	*+1.37%*	52.86%	*+2.48%*	50.10%	*+1.42%*
		BestConcept	63.00%	-2.78%	45.80%	-2.14%	64.30%	-2.13%	51.40%	*+0.19%*	25.70%	-12.29%	50.04%	-2.99%	47.30%	-4.25%
		BestConceptID	57.50%	-11.27%	38.80%	-17.09%	60.70%	-7.61%	45.80%	-10.72%	26.50%	-9.56%	45.86%	-11.09%	44.30%	-10.32%
	Complete	AllConcepts	64.60%	-0.31%	49.60%	*+5.98%*	67.40%	*+2.59%*	52.90%	*+3.12%*	28.00%	-4.44%	52.50%	*+1.78%*	49.70%	*+0.61%*
		AllConceptsID	63.30%	-2.31%	51.10%	*+9.19%*	67.70%	*+3.04%*	53.00%	*+3.31%*	29.40%	*+0.34%*	52.90%	*+2.56%*	50.10%	*+1.42%*
		BestConcept	63.50%	-2.01%	45.80%	-2.14%	65.70%	+0.00%	50.90%	-0.78%	25.00%	-14.68%	50.18%	-2.71%	47.40%	-4.05%
		BestConceptID	59.00%	-8.95%	39.70%	-15.17%	62.60%	-4.72%	47.20%	-7.99%	25.80%	-11.95%	46.86%	-9.15%	45.00%	-8.91%

As illustrated in the table, in most cases the original system outperforms the new system integrating conceptualization phase. In fact this applies to approximately 60% of results evaluated using the micro-average F-measure. However, after a thorough look into the results, it seems clear that the system using the similarity measure of KullbackLeibler and NB shows some amelioration. Indeed, results using Micro averaged F-measure are improved in two thirds of the cases after conceptualization.

Considering each classes independently in the results, we can observe that conceptualization improves the outcome in about 58% of cases for the class "C23". All methods, except for SVMs, when tested on the original corpus, showed the worst performance in treating this class. This difference is significant according to the McNemar [20] test on classification results considering C23's documents before and after conceptualization ($\rho << 0.01$). Moreover, we observe a significant improvement of the outcome in about 66% of cases for the class "C06" when using SVMs.

Concerning conceptualization strategies, one of the 12 conceptualization strategies tested in these experimentations seems to provide an improvement in most cases. Indeed, Add Best Concept strategy outperforms others as it retrieves the best candidate concept provided by MetaMap as the mapping result. Best concept names rather than IDs are used in text conceptualization. However this strategy does not present a significant improvement over the original system without conceptualization. Indeed, the gain is in most cases around 1% or less. The highest increases are obtained when using conceptualization strategies taking into account all the candidates found by MetaMap and not only the best. Thus, in some cases this increase exceeds 60%. Furthermore, largest increase in the Micro-average F-Measure values are attained using the strategy of adding Concepts applied to the NB. Previous improvements seem to be significant according to McNemar [20] test having ($\rho << 0.01$).

Concerning Rocchio method used with different similarity measures, the least improvement of conceptualization can be observed for the method using the similarity measure Cosine. In fact, among all Micro-averaged F-measure values for this method, only one surpasses its value when compared to system results using Cosine without conceptualization. Rocchio using Jaccard similarity measure outperforms the original system in conceptualizing text using the best concept name according to three different strategies: (AddConcepts, Partial and Complete) conceptualization. Considering Rocchio with Kullback, the improvement is obtained in the case of class “C23” with all strategies (Partial, Addconcept, or Complete). Moreover, only the strategy Add Bestconcept improves the original method in four of five cases.

Concerning SVMs is improved in most cases when using *AllConceptID* regardless the strategy (AddConcept, Partial or Complete).

Considering NB, an improvement is observed in almost all cases when the strategy is *AddConcept*. This improvement occurs in approximately 58% of cases.

4.4 Discussion

According to the results presented in the preceding section, here we list some remarks. First of all, in most case, lowest results are observed when terms are replaced by IDs of their corresponding concepts in the UMLS. This performance degradation might be principally related to replacing all terms corresponding to a concept by its ID; only the IDs of concepts can participate in vectorization. Terms that are shared among concept with different IDs are excluded from vectors even if they had a high importance.

Second, when the system performance has a good F1-measure value (i.e. exceeds 60%), no significant effect can be observed for the integration of the conceptualization task into the system. In fact, as the same similarity measures are used for both cases with/without conceptualization, results' amelioration was limited.

Third, when the system performance using a specific similarity measure has a low F1-measure value, as it the case for the class "C23" or "C06" in the case of SVMs, introducing conceptualization can significantly improve this value with a maximum gain reaching (60%) in some cases. Indeed, the class "C23" is very large compared to others and so enriching class representation by semantics might result in a better identification of this class and also in better results.

Fourth, the best conceptualization strategy is Addconcept adding the Best concept among mapped candidates into the text. In fact, best mappings retrieved by MetaMap are added into text in order to enrich it with semantics avoiding any information loss.

Finally, even if the results are still preliminary, it seems useful to introduce semantic enrichments to classification methods in order to ameliorate their predictions. However, these improvements are relatively dependent on the behavior of the method and also on used corpus and its class distribution. Consequently, it seems necessary to experimentally define the conditions under which the introduction of semantics can improve classification. Moreover, the exploitation of semantic resources was limited in this work. For example, it ignores all relations (like Subsumption and Transversal relations) among concepts used in the conceptualization task. Thus, it seems adequate to deploy these relations in the classification process.

5 Related Work

Text classification is a challenging task due to the sparse and high dimensional feature space. Moreover, the complex nature of semantics residing in text makes text classification so difficult due to ambiguities that are tricky to resolve. Most of these difficulties are related to the widely used text representation model (VSM) that is unable to extract useful and meaningful features from text.

Considering Kernel based classifiers like SVMs, many works define kernel function on the knowledge base hierarchy producing semantic kernels. Séaghdha [21] proposes multiple semantic kernel functions on WordNet and proves that SVMs work better with semantic kernels.

Many works proposed new extensions to the traditional VSM in order to overcome its limitations. Numerous weighting schemes for the traditional VSM are proposed in [22], all aiming at optimizing feature weights, which might ameliorate text classification. Moreover, other works demonstrate some improvements by aide of new feature extraction methods. In order to overcome the VSM's limitation that is related to composed words, authors in [23] propose a Bag of Phrases (BOP) model instead of the traditional Bag of Words (BOW) taking frequently occurring N-gram phrases into account during feature extraction. According to tests on KNN, Decision Trees, SVMs and NB, the proposed BOP outperforms the original BOW. Despite the improvements demonstrated in this work, few of VSM limitations are treated.

In order to take into account ambiguities and polysemous words beside the previous limitations, background knowledge bases are frequently used. Indeed, thesaurus and domain specific ontologies can help in determining the accurate semantics residing in text. According to these approaches, the original BOW is transformed into Bag of Concepts (BOC) [24] so resulting vectors are constituted of text-related concepts. These concepts can be discovered in the original text by means of background knowledge. It is also important to incorporate related concepts to those already adopted through conceptualization step. This deployment enriches document representation and might improve classification results.

Authors in [24] present a new representation model using concepts extracted from background knowledge. AdaBoost algorithm is tested on three different corpora in order to support the approach. During the experimentations on the corpus Reuters-21578, Wordnet is used as the background knowledge whereas MESH ontology is used with OHSUMED dataset. Different strategies of sense disambiguation were used. Moreover, the superconcepts of specific concepts discovered in text are also integrated into the vector of concept representing text documents. These superconcepts are searched up to a maximal distance into the ontology. Deploying superconcept in document representation model is called generalization which helps in improving results especially with general purpose ontologies like WordNet. Nevertheless, this strategy does not seem to be adequate when dealing with domain specific knowledge bases like MESH.

Bai, Wang [25] align three general purpose ontologies: WordNet, OpenCyc and SUMO in their system and use the resulting knowledge base. The traditional BOW model is then replaced by BOC through new ontological indexing of text documents by means of this knowledge base. The context strategy is used in order to resolve ambiguities. Text classification using SVMS is realized on three corpora: Reuters-21578, OHSUMED and 20Newsgroups. Significant improvements are demonstrated especially with OHSUMED data set. Authors conclude that integrated concepts are helpful for special domain datasets.

Guisse, Khelif [26] propose also a BOC approach with generalization using weight propagating algorithm in order to attribute appropriate weights to superconcepts in the domain specific ontology. This work demonstrated significant improvement in patent classification.

New representation models using parts of ontology hierarchy are also proposed in the literature. These parts constitute semantic trees [27] or forests [28] where each concept is assigned an importance score. Using the semantic hierarchy of WordNet, significant improvement is demonstrated in classifying Yahoo! document [27]. Concept Forests [28], that are constituted of parts of WordNet, help in improving classification results when tested on Reuters-21578.

According to the literature, Authors seem to disagree in assessing the importance of using semantic information in classification [29]. Nevertheless, it seems to be a promising approach taking into consideration the particular context of the classification task [6].

6 Conclusion and Perspectives

Due to the explosive growth of published data, many search engines demonstrate poor performance that cannot meet the needs of users. This leads to a challenging need for efficient filtering, ranking and classification techniques. This paper concerns text classification that is currently a challenging research topic, particularly in domains as information retrieval, recommendation, personalization, user profiles etc.

Generally, text classification uses syntactical and statistical models for text document representation as it is the case for traditional methods as Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio. Representations resulting from these models suffer the lack of sense ignoring all semantics that reside in the original text that can help in text classification. Vector-based representation used in these methods permits the integration of semantics through "Conceptualization" that enriches document representation.

This work, depending on experiments, tries to estimate the impact of different text conceptualization strategies on traditional classification methods. These experiments are realized in the biomedical domain and particularly on the OHSUMED corpus, using domain specific knowledge base UMLS. Three traditional classification methods are chosen for these experiments: Rocchio (using different similarity measures), SVMs and NB. Indeed, the results of our experiments show, in some cases, considerable performance improvements. We also observed that these improvements are related to many factors: the classification method, the corpus, the class and the conceptualization strategy.

Finally, it seems useful to integrate semantic knowledge into text representation in order to ameliorate text classification. Nevertheless, experimental analysis is necessary to determine if this integration is appropriate and in which conditions it might be applied. A combination of the original and the enriched representation model can be used in order to combine their advantages and overcome their limits. In addition, other useful information contained in semantic resources such as relations can also be deployed in decision making leading to semantic classification that can improve search in restricted domains.

References

1. Asirvatham, A.P., Ravi, K.K.: Web page classification based on document structure (2001)
2. Ferreira, R., et al.: Improving News Web Page Classification Through Content Extraction. In: IADIS International Conference WWW/Internet 2011 (2011)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Computer. Survey 34(1), 1–47 (2002)
4. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
5. Hotho, A., Staab, S., Stumme, G.: Text clustering based on background knowledge (2003)
6. Ferretti, E., Errecalde, M., Rosso, P.: Does Semantic Information Help in the Text Categorization Task? Journal of Intelligent Systems 17, 91–107 (2008)
7. Garla, V.N., Brandt, C.: Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. (in press)
8. Yetisgen-Yildiz, M., Pratt, W.: The effect of feature representation on MEDLINE document classification. In: AMIA Annu. Symp., pp. 849–853 (2005)

9. Zhang, X., Jing, L., Hu, X., Ng, M., Zhou, X.: A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 115–126. Springer, Heidelberg (2007)
10. Camous, F., Blott, S., Smeaton, A.F.: Ontology-Based MEDLINE Document Classification. In: Hochreiter, S., Wagner, R. (eds.) BIRD 2007. LNCS (LNBI), vol. 4414, pp. 439–452. Springer, Heidelberg (2007)
11. Unified Medical Language System (UMLS®), `http://www.nlm.nih.gov/research/umls/`
12. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. J. Am Med. Inform. Assoc. 17(3), 229–236 (2010)
13. Hersh, W., et al.: OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 192–201. Springer-Verlag New York, Inc., Dublin (1994)
14. Medical Subject Headings (MeSH®), `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`
15. Yi, K., Beheshti, J.: A hidden Markov model-based text classification of medical documents. J. Inf. Sci 35(1), 67–81 (2009)
16. Huang, A.: Similarity measures for text document clustering. In: Sixth New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, pp. 49–56 (2008)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 1–27 (2011)
18. Hall, M., et al.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
19. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing Management 45(4), 427–437 (2009)
20. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10(7), 1895–1923 (1998)
21. Séaghdha, D.O.: Semantic classification with WordNet kernels. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 237–240. Association for Computational Linguistics, Boulder (2009)
22. Lan, M., et al.: Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Trans. Pattern Anal. Mach. Intell. 31(4), 721–735 (2009)
23. Li, Z., Li, P., Wei, W., Liu, H., He, J., Liu, T., Du, X.: AutoPCS: A Phrase-Based Text Categorization System for Similar Texts. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, Q.-M., et al. (eds.) APWeb/WAIM 2009. LNCS, vol. 5446, pp. 369–380. Springer, Heidelberg (2009)
24. Bloehdorn, S., Hotho, A.: Boosting for Text Classification with Semantic Features. In: Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (eds.) WebKDD 2004. LNCS (LNAI), vol. 3932, pp. 149–166. Springer, Heidelberg (2006)
25. Bai, R., Wang, X., Liao, J.: Using an Integrated Ontology Database to Categorize Web Pages. In: Kim, T.-H., Adeli, H. (eds.) AST/UCMA/ISA/ACN 2010. LNCS, vol. 6059, pp. 300–309. Springer, Heidelberg (2010)
26. Guisse, A., Khelif, K., Collard, M.: PatClust: une plateforme pour la classification sémantique des brevets. In: Conférence d'Ingénierie des connaissances, Hammamet, Tunisie (2009)

27. Peng, X., Choi, B.: Document classifications based on word semantic hierarchies. In: International Conference on Artificial Intelligence and Applications (AIA 2005), pp. 362–367 (2005)
28. Wang, J.Z., Taylor, W.: Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2007, pp. 395–401. IEEE Computer Society (2006)
29. Stein, B., Eissen, S.M.Z., Potthast, M.: Syntax versus semantics: Analysis of enriched vector space models. In: Third International Workshop on Text-Based Information Retrieval (TIR 2006). University of Trento, Italy (2006)