Concept-based Topic Model Improvement

Claudiu Musat[1], Julien Velcin[2], Marian-Andrei Rizoiu[2] and Stefan Trausan-Matu[1],

[1] „Politehnica" University of Bucharest, 313 Splaiul Independentei,
Bucharest, 060032, Romania
{claudiu.musat, trausan}@cs.pub.ro
[2] Laboratoire ERIC, Universite Lyon 2, 5 Avenue P. Mendes,
Bron, 69676, France
{marian-andrei.rizoiu, julien.velcin}@univ-lyon2.fr

Abstract. We propose a system which employs conceptual knowledge to improve topic models by removing unrelated words from the simplified topic description. We use WordNet to detect which topical words are not conceptually similar to the others and then test our assumptions against human judgment. Results obtained on two different corpora in different test conditions show that the words detected as unrelated had a much greater probability than the others to be chosen by human evaluators as not being part of the topic at all. We prove that there is a strong correlation between the said probability and an automatically calculated topical fitness and we discuss the variation of the correlation depending on the method and data used.

Keywords: Topic Models, Ontologies, Evaluation, Improvement

1 Introduction

A decade ago, when dealing with textual datasets, one usually had to choose between a linguistic approach and a statistical one, between going in depth with semantic issues using thesauri and extra-knowledge and being able to process large amounts of data. Today a main research direction is to build models that benefit from both worlds [1]. In particular, the stake lies mainly in embedding syntax and semantics into powerful statistical models [2].

Topic models are Bayesian statistical models that have proven their accuracy in many applicative contexts [3]. Given a large corpus, these models permit the extraction of topics that structure the texts and the topics themselves can be simplified to a list of keywords. However, using only statistical properties is clearly not enough to obtain good topics. Stopwords and outliers often pollute the topics and make their meaning obscure. Using external knowledge, such as an ontology, is a privileged track to improve the topic quality. Until now, this research track has been little explored [4].

In this paper, we propose the use of WordNet [5] as a post-processing step for detecting and removing outliers from the topic labels. However, any concept hierarchy usually found in domain ontologies can be used. The idea is to create a projection of the topic as a whole onto the given ontology and decide which part of the topic – if any – is separated from the others. We can improve the

understandability of the given topic if we are able to remove its parts that are unrelated from a human perspective.

We performed multiple experiments on two different corpora: a general dataset on American history and the second is a specific dataset containing exclusively economic articles. We asked 37 external and independent humans to judge the quality of our model's outputs. The results show clearly that the algorithm follows human intuition and that improving topics in this manner is feasible. The paper continues with an outline of the state of the art in section 2, the proposed method is detailed in section 3, while results and conclusions follow in sections 4 and 5 respectively.

2 State of the art

Literature proposes multiple ways of how to extract meaning from text. Approaches coming from the field of Natural Language Processing (NLP) [6] start from the analysis of the text (using, for example, a Part-Of-Speech tagger) and only then use statistical information to ameliorate the result. Others [7] employ methods inspired from clustering, by first translating the documents into the space-vector model.

In recent years, generative methods imposed as the state-of-the-art for their proven results. Latent Dirichlet Allocation (LDA) [8] is a probabilistic generative model designed to extract topics from text corpora. It starts from the bag-of-words model that considers documents as collections of words, without making use of their order. The presence of each word is considered to be generated by hidden latent variables, the topics. Each document is represented as a list of mixing proportions of the topics.

Generative methods like LDA are purely statistical approaches, that only consider information such as the number of appearances of words into documents. While the complex mathematical model constructed in order to approximate the hidden generative variables (the topics) does succeed to catch some of the meaning of the texts, even dealing with problems like polysemy, there is room for improvement.

2.1 Topic Evaluation and Improvement

Topic models have been evaluated in both a quantitative and a qualitative manner. Qualitatively, a topic is represented by a short list of words in order to convince the reader of their usefulness and either the user or the author usually attach a label to it. Quantitatively the perplexity measure [9] has been one of the most widely used metrics. However it has been shown [10] that human judgment does not always coincide with these common evaluation criteria. This finding has prompted other researchers to look for novel evaluation systems such as that proposed in [11].

Model improvement ideas varied from supervision [12] in the topic generation to considering semantic information. Wang *et al.* [13] go beyond the bag-of-words approach and devise a generative topic model that is based on n-grams, instead of words. Other works induced a correlation structure between the constructed topics [14]. External resources were also used. WordNet [5], for example, has a long tradition of being used in text classification tasks [15]. In [4], WordNet is used

together with a generative topic model for word sense disambiguation. Starting from the idea of the WordNet-Walk algorithm, word senses are used as a hidden variable.

Topic Detection and Tracking (TDT) is a domain that uses topic modeling algorithms extensively for multiple tasks (e.g. modeling topics through time [16] or extracting topic trees [17]). Works like [18] incorporate semantic information in the language modeling framework. Other semantic resources, like places, dates, names, are used to delimit the time intervals and evolve topics. Information Retrieval is another domain that benefits from using semantic information. Mihalcea and Mihalcea [19] show that retrieval effectiveness can be improved by indexing words with their semantic classes such as parts-of-speech, named-entity-type, WordNet synonyms, hypernyms, hyponyms, etc.

Latent Dirichlet Allocation with WordNet [20] is a version of LDA that uses the word senses as hidden variables and tries to select the right sense when constructing topics, becoming a tool for word sense disambiguation. WordNet is one of the most used resources for sense disambiguation and several ways of using it are discussed by Navigli [20]. The sense disambiguation techniques for detecting words which are semantically related may be used for post processing topic models.

2.4 WordNet

WordNet may be considered a general ontology or a lexical database. It is in fact a huge semantic network linking the majority of usual words in English through a fixed set of relations like: synonymy, hypernomy/hyponymy (super/sub concept), meronymy/holonymy (part/whole), antonymy, etc. Each word may have several senses and for each sense it has a set of synonyms (a synset). Each synset represent a distinct concept and semantic distances between pairs of words may be computed [21]. Consequently, sets of words may be grouped in semantic neighborhoods.

There are several differences between WordNet semantic neighborhoods and semantic spaces of LSA or topics discovered with LDA. First of all, the former are obtained from the word networks built explicitly by humans, starting from psycholinguistics data, while LSA and LDA word grouping is determined statistically from text corpora. The advantage of using WordNet is precision while the disadvantage is the lack of dynamics and of the possibility to handle very specific domains. Even if it has more than 200,000 word-sense pairs, WordNet cannot cope with very specialized terms or neologisms. A second difference is that in WordNet words are not only grouped by similarity, they are also related by various relations, as mentioned above and thirdly, each word in WordNet has a gloss. The latter two features may be exploited for further semantic processing.

3 Proposed System

The presented system is designed to improve individual topics according to their conceptual cohesion. We use an established [10] simplified representation of each topic within a model, a list of its top words. To improve topic readability and meaningfulness we prune the topical top words that are unrelated to the others from a

conceptual perspective. The remaining ones are thus more inter-related as a set and confer more meaning to the user. The kernel of this work is establishing the conceptual context of a single given topic. We detect which concepts from the used ontology are relevant to the topic as a whole and output the topical words unrelated to those concepts as the outliers to be eliminated.

3.1 Structure Representation

A definition of topic models states that they are sets of discrete probability functions z over a given text collection T, $\{p(w|z)\}_{w \in z}$, with w being words from the employed vocabulary V, $w \in V$. Each topic z is one of the k topics in the model obtained using a known algorithm given T, $\Theta = \{z_1, \dots, z_k\}$. We reduce the representation of each topic z to the set of its most relevant words, the ones with the highest probability given z, $\omega(z) = \{w_{1,z}, w_{1,z}, \dots, w_{n,z}\}, n \in \mathbb{N}$.

Prior knowledge about the related concepts is structured in an ontology $O(D, \mathcal{R})$, where D is the set of all concepts and $\mathcal{R}$ the set of all possible relations within O. We assume there exists a relation $r \in \mathcal{R}$ according to which all concepts in a subset $D_r \subset D$ form a tree $\mathcal{C}(D_r, r)$ in which the concepts in D_r are the $\mathcal{C}$ nodes or $\delta(\mathcal{C})$ and r is the relation between them. Possible examples include the hypernymy or hyponymy relations between concepts such as those in WordNet [5]. We further assume that for a subset of words in the vocabulary $V_\mathcal{C} \subset V$ there is a non void at least one concept in $\mathcal{C}$ that is a sense of a given word $w \in V_\mathcal{C}$. Given w, let $\delta(w) \subset \delta(\mathcal{C})$ be the set of all senses of the word w within $\mathcal{C}$.

3.2 Ontological Subtrees

We use the *ancestral path* distance to determine the distance between two concepts, $d(c_i, c_j)$ with $c_i, c_j \in \mathcal{C}$. The distance is infinity if one is not an ancestor of the other. We define a *branch* as a path between two concepts c_i and c_j where one is either a direct or indirect ancestor of the other.. Let $\delta(c_i, c_j)$ be the set of nodes located on the branch that connects c_i and c_j. Let $c_0(\mathcal{C}) \in \delta(\mathcal{C})$ be the root of the $\mathcal{C}$ tree. Let the *distance* between a word $w \in V_\mathcal{C}$ and a concept c $\in \mathcal{C}$, $dist(w, c)$, be the minimum of all the distances between the word's senses $c_w \in \delta(w)$ and the target concept.

Definition 1. The subtree of an arbitrary concept c within $\mathcal{C}$ is the subtree of $\mathcal{C}$ whose root concept is c, . $\mathcal{C}_{Metallic\_element_1}$ is presented in Fig. 1 in a rectangle.

$$\mathcal{C}_c = (\delta(\mathcal{C}_c), r); \ \delta(\mathcal{C}_c) \subset \delta(\mathcal{C}), \delta(\mathcal{C}_c) = \{c_i \in \delta(\mathcal{C})| c_i \xrightarrow{r^*} c\} . \tag{1}$$

Definition 2. A word's w subtree of a concept c within $\mathcal{C}$, $\mathcal{C}_{w,c}$ is the subtree of $\mathcal{C}_c$ that contains all the branches between concepts within $\delta(w)$ and subtree the root c. $\mathcal{C}_{w_2,metallic\_element\#1}$ is in Fig 1. a reunion of the $(gold\#3, noble_{metal}\#1)$ and $(noble_{metal}\#1, metallic_{element}\#1)$ arcs.

$$C_{w,c} = (\delta(w, \mathcal{C}_c), r); \; \delta(w, \mathcal{C}_c) = \bigcup_{c_w \in \delta(w)} \delta(c_w, c) \;. \tag{2}$$

Given a topic z_i and its most important words, $\omega(z_i)$, let a topic's relevant concepts $\delta(z_i) \subset \delta(\mathcal{C})$ be the reunion of the $\mathcal{C}$ senses of the topic's relevant words.

Definition 3. A topical subtree of a concept c within $\mathcal{C}$ is the reunion of all the $C_{w,c}$ subtrees of all topical words ($C_{z_i, abstraction\#6}$ is shown in the Fig 1. with a bold line):

$$C_{z_i,c} = (\delta(z_i, \mathcal{C}_c), r); \; \delta(z_i, \mathcal{C}_c) = \bigcup_{w \in \omega(z_i)} \bigcup_{c_w \in \delta(w)} \delta(c_w, c) \;, \forall z_i \in \Theta \;. \tag{3}$$
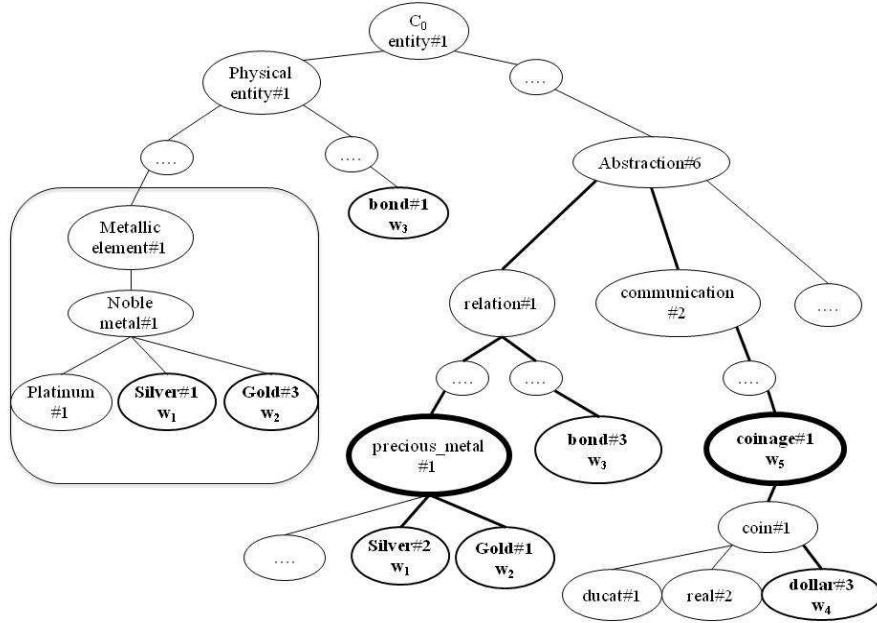

Fig 1. The WordNet subtree of *metallic_element#1* lies within the *rectangle's* borders, the topical subtree of *abstraction#1* is shown with a *bolded line*, concepts which are senses of topic words are *bolded* and relevant concepts to the topic have a *heavily bolded contour*.

3.3 Concept Relevance and Topical Outliers

In order to detect the topical words that are unrelated to conceptual context created by the others we must first identify the related concepts. We aim to detect the topical subtrees that include as many of the topic's words as possible (at least one sense for each) while at the same time having a root concept as specific as possible. Specificity in this case is determined by the node's height – its distance to the ontology root $c_0(\mathcal{C})$ and its depth – how far it is from the subtree leaves. The greater its height, the more specific the concept is, while the greater its depth, the more general it becomes.

Definition 4. A concept's relevance $\phi: \mathcal{C}_{z_i} \times \Theta \to \mathbb{R}_+$ to a given topic is a weighted average of its coverage $cov: \delta(z_i, \mathcal{C}) \times \Theta \to \mathbb{N}_+$, height $h: \delta(\mathcal{C}) \to \mathbb{N}_+$ and depth $\rho: \delta(z_i, \mathcal{C}), \Theta \to \mathbb{N}_+$, with the weights ω_{cov}, ω_h and ω_ρ respectively.

$$c \mapsto \phi(c) = \omega_{cov} \cdot \frac{card\{\cap\, \delta(z_i), \delta(z_i, \mathcal{C}_c)\}}{card\{\delta(z_i)\}} + \omega_h \cdot d\big(c, c_0(\mathcal{C})\big) - \omega_\rho \tag{4}$$
$$\cdot \text{avg}\big[\big(d(w, c) | c \in \delta(z_i)\big)\big].$$

The higher the relevance of the concept with the highest fitness value, the higher the topic's cohesion viewed as a whole. But aside from the evaluation function of the concept relevance assessment, one can also improve the initial model based on the said assessment, through the detection and elimination of conceptually outlying words, or topical outliers.

Definition 5. Topical outliers $w_i \in \omega_O(z_i) \subset \omega(z_i)$ are words not covered by the reunion of the topical subtrees of the concepts with the highest $l \in \mathbb{N}$ topical relevance values given z_i, $c_{z_i}^*$, with l an experimentally established parameter:

$$\omega_O(z_i) = \omega(z_i) - \bigcup_{c_{z_i}^*} \delta\left(\mathcal{C}_{c_{z_i}^*}\right). \tag{5}$$

In Fig. 1 we present a simplified version of the topical subtree for the topic $z = (silver, gold, bond, dollar, coinage) = (w_i); 1 \le i \le 5; i \in \mathbb{N}$, extracted using LDA from the Suall dataset [13]. Due to space considerations, not all word senses are shown. The WordNet concepts are shown in a $word\#sense$ format. The ones with a highest calculated relevance that have a *distinct* topical coverage given z were, in a decreasing order of relevance, $\delta(precious\_metal\#1) = \{silver, gold\}$, and $\delta(coinage\#1) = \{dollar, coinage\}$, while *bond* was the obtained outlier. The two concepts are outlined with a heavily bolded contour in Fig.1. Because $noble\_metal\#1$ is higher in the WordNet hierarchy than $precious\_metal\#1$ (the difference between the two being the chemical or financial standpoint), although they both have a coverage of 2 and a depth of 1, the latter was chosen to represent w_1 and w_2. The third sense of w_4, dollar#3 is a hyponym of the first sense of $coinage$, while w_4's ancestor with the best topical relevance is $communication\#2$, a very general concept which makes w_4, $bond$, an outlier.

4 Experiments

We obtained 10 different topic sets by running the LDA algorithm built into the Mallet suite [22] with five different k values $k \in \{30, 50, 100, 200, 300\}$ on two corpora. We chose two corpora to find whether results would differ greatly from a general purpose corpus such as the Suall [13] to a targeted one, in our case an economic corpus. The second corpus contains 23986 publicly available Associated Press articles published in the Yahoo! Finance section between July and October 2010.

We benchmarked the results of our outlier detection algorithm with human evaluations similar to those employed by [10]. Evaluators were asked to extract the unrelated words from a group containing the top five words from one topic and an additional spurious word. One or more unrelated words were chosen for each group. We test whether topic words that were marked as outliers by our algorithm have a better than average chance of being wrongly marked as the spurious word that is inserted within the topic.

The choice of the spurious word is not obvious as it greatly influences the outcome of the experiment. While Chang *et al.* [10] use a *random* word from those irrelevant to the current one, we discuss two opposing scenarios. Within each model, all the inter-topic Kullback – Leibler (KL) divergences are computed and for each topic we determine which topics are closest and farthest. We then randomly select a word from the top five from both the closest and the farthest topics which will be used further as spurious words. For instance, given the above ($silver, gold, bond, dollar, coinage$), the word chosen from the closest KL neighbor was *specie* while the choice from the farthest topic was *technology*. When the latter is mixed with the five original topic words, it is to spot as the real spurious word, which makes it harder to also detect *bond* as unrelated.

4.1 Experiment Framework

The experiments below were devised to answer two questions – are topical outliers more likely to be marked as spurious words by human evaluators? If so, what does the probability of this happening depend on? A total of 37 evaluators were each given 40 groups of six words in a random order containing five topic words and one spurious word. The questions were balanced to have an equal number of topics evaluated for the two corpora – Suall and the economic AP – for each topic number k and for each of the two spurious word types. From each experiment, only the top and bottom ten topics were considered, ordered by the fitness of their representative concept. Outliers were algorithmically outputted if they were not covered by the most important two concepts related to the topic. Not all topics had detectable outliers and they were removed from the experiment, thus the total number of topics in each run varies. Detection results are shown separately for the best and worst topics in each case.

4.2 Discussion

We compared the probability of an algorithmically calculated outlier being marked as a spurious word by the evaluators with the probability of a regular non-outlier word being marked. In Table 1, *total* represents the total number of considered topics for that particular situation, from which in *hit* cases the spurious word was detected, while in o*ut* cases in the spurious word was not hit, but an outlier was. The odds for the outlier to be hit marked are given by *p(out)* while the probability for a regular word to be marked as spurious is *p(other)*.

Table 1. Outlier detection and correlation for the worst topics obtained from the Suall corpus

Topic	Corpus	Spurious word	K	Total	Hit	Out	p(out)	p(other)	Gain	Fit	
Top 10	Suall	Close	30	14	3	8	0.73	0.07	1067%	2.22	
			50	26	3	3	0.13	0.22	60%	2.08	
			100	33	7	7	0.27	0.18	147%	2.29	
			200	25	11	3	0.21	0.20	109%	2.30	
			300	21	5	9	0.56	0.11	514%	2.40	
		Pearson									0.14
		Distant	30	9	5	3	0.75	0.06	1200%	2.30	
			50	30	14	2	0.13	0.22	57%	2.08	
			100	34	19	4	0.27	0.18	145%	2.29	
			200	15	10	1	0.20	0.20	100%	2.33	
			300	12	7	2	0.40	0.15	267%	2.34	
		Pearson									0.3
Top 10	AP	Close	30	8	3	0	0	0.25	0%	1.54	
			50	16	4	4	0.33	0.17	200%	2.25	
			100	11	3	3	0.38	0.16	240%	2.38	
			200	16	8	3	0.38	0.16	240%	2.18	
			300	25	12	6	0.46	0.13	343%	2.32	
		Pearson									0.91
		Distant	30	9	2	1	0.14	0.21	67%	1.54	
			50	17	6	3	0.27	0.18	150%	2.25	
			100	12	8	2	0.50	0.13	400%	2.38	
			200	17	8	3	0.33	0.17	200%	2.12	
			300	8	5	1	0.33	0.17	200%	2.08	
		Pearson									0.78
Bottom 10	Suall	Close	30	25	6	11	0.58	0.11	550%	1.50	
			50	41	10	9	0.29	0.18	164%	1.32	
			100	25	3	2	0.09	0.23	40%	1.17	
			200	31	2	6	0.21	0.20	104%	1.05	
			300	11	1	0	-	0.25	0%	1.16	
		Pearson									0.87
		Distant	30	27	10	8	0.47	0.13	356%	1.50	
			50	42	16	9	0.35	0.16	212%	1.32	
			100	19	8	2	0.18	0.20	89%	1.23	
			200	35	17	4	0.22	0.19	114%	1.05	
			300	11	6	2	0.40	0.15	267%	1.16	
		Pearson									0.72
Bottom 10	AP	Close	30	18	7	5	0.45	0.14	333%	1.16	
			50	10	5	1	0.20	0.20	100%	1.05	
			100	23	5	0	-	0.25	0%	1.15	
			200	8	4	0	-	0.25	0%	0.97	
		Pearson									0.21
		Distant	30	13	5	3	0.38	0.16	240%	1.17	
			50	11	7	1	0.25	0.19	133%	1.05	
			100	31	13	2	0.11	0.22	50%	1.10	
			200	3	2	0	-	0.25	0%	0.87	
			300	6	3	3	1.00	-	**1000%**	1.09	
		Pearson									0.33

We prove that outliers do have a significantly larger probability of being marked as spurious words than other words, on average 238% more for the bottom ranking topics and 285% for the best ones. While this demonstrates that algorithmically detected outliers are likely to be viewed as such by a human mind as well, we are still interested in finding the correlation between the human outlier detection rate (as expressed by the probability gain) and topical relevance, given by the value for its most important concept.

We compute the Pearson correlation between the two and show it in dedicated rows in Table 1. A value close to 1 or -1 implies strong positive or negative correlation while values close to 0 show a lack of linear correlation. We observe the correlation between the probability gain and the topical fitnesses varies from medium to very strong positive values and that it depends on the corpus. The economic corpus is predictable only for the good topics while Suall is easy to improve for its lower end, probably because its worst topics are much better (higher fitness) than the economic bad topics. Also, correlation depends on the way the spurious word is chosen. It is always more extreme for the poisoning with words from similar topics. A very similar spurious word coupled with a good topic immediately reveals the outlier; coupled with a bad one only adds to the general confusion.

4 Conclusions and Future Work

We have proposed and successfully tested a hypothesis in which conceptual knowledge used in a post processing phase improves topic model output by removing unrelated words from the simplified topic description. The improvement is in line with human judgment, a fact proven by the correlation between automatically obtained results and human outlier detection rate. Also, topics obtained from the economic corpus are more understandable than their peers drawn from the Suall set – a rather counter intuitive finding, given the expectation that WordNet would better portray the concepts behind the topics drawn from a more general dataset.

It is noteworthy that although the method was tested using WordNet and LDA, it can be easily extended to other ontologies and topic modeling algorithms. Future work includes a test framework that can compare multiple topic models and use WordNet for languages other than English. We also plan to quantify the role of the context created by other topics when analyzing a single one and in the nearest future to label topics from a conceptual standpoint rather than a statistical one.

References

1.   Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT

Press, Cambridge (1999)

2. Boyd-Graber, J., Blei, D.M.: Syntactic Topic Models. Journal Computational Linguistics (2008)

3. Blei, D.M., Lafferty, J.: Text Mining: Theory and Applications, chapter Topic Models. Taylor and Francis, London (2009)

4. Boyd-Graber, J., Blei, D.M., Zhu. X.: A topic model for word sense disambiguation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1024--1033 (2007)

5. Miller, G.A.: WordNet: a lexical database for English. Journal Communications of the ACM. 38(11), pp. 39-41, ACM (1995).

6. Hammouda, K.M., Matute, D.N., Kamel, M.S.: Corephrase: Keyphrase extraction for document clustering. J. MLDM, pp. 265–274 (2005).

7. Rizoiu, M.-A., Velcin, J. and Chauchat, J.-H. Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. 10éme conférence Extraction et Gestion des Connaissances (EGC 2010), Hammamet, Tunisie, 2010.

8. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3, pp. 993 – 1022 (2003)

9. Wallach, H.M., Murray, I., Salakhutdinovn, R., Mimno, D.: Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1105 – 1112 (2009)

10. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models. Journal Neural Information Processing Systems. 31 (2009)

11. Newman, D., Lau, J.H., Grieser, K., Baldwin,T.: Automatic Evaluation of Topic Coherence. In The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100 – 108 (2010)

12. Blei, D.M., McAuliffe, J.: Supervised topic models. Advances in Neural Information Processing Systems. 20, pp. 121—128 (2008).

13. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th ieee International Conference on Data Mining, pp. 697–702 (2007)

14. Lafferty, J.D., Blei, M.D.: Correlated topic models. In Advances in Neural Information Processing Systems, Proceedings of the 2005 conference. pp. 147 – 155 (2006)

15. Scott, S., Matwin, S.: Text Classification using WordNet Hypernyms. In Proceedings of the Association for Computational Linguistics Conference. pp. 38 – 44 (1998)

16. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In Proceedings of the 23rd international conference on machine learning. pp. 120 – 128 (2006).

17. Blei, D.M., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems, 16 , 106 – 114 (2004).

18. Nallapati, R.: Semantic language models for topic detection and tracking. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop. vol. 3, Association for Computational Linguistics, pp. 1 – 6 (2003)

19. Mihalcea, R.F., Mihalcea, S.I.: Word semantics for information retrieval: moving one step closer to the Semantic Web. In: Tools with Artificial Intelligence, Proceedings of the 13th International Conference on. IEEE, pp. 280 – 287 (2002)

20. Navigli, R.: Word Sense Disambiguation: A Survey,ACM Computing Surveys, Vol. 41, No. 2, Article 10 (2009)

21. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. Computational Linguistics, 32(1), pp. 13—47 (2006).

22. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. (2002)