Automatic Identification of Domain Terms: An Approach for Italian

Maria Teresa Artese, Isabella Gagliardi[0000-0002-4667-919X]

IMATI – CNR, Via Bassini 15, 20133, Milan, Italy
artese@mi.imati.cnr.it, gagliardi@mi.imati.cnr.it

Abstract. The problem of creating a fully automated specific-domain thesaurus is very topical. The paper presents a novel method to address this problem in the Italian language. The main feature of this approach is the integration of different methods: machine learning classification methods working on the semantic representation of candidate terms, word embeddings models, able to capture the semantics of words, and a computation of the degree of specialization of a term. The work is in progress and results obtained so far are promising.

Keywords: Classification Methods, Word Embedding Models, Probability, Food, Italian Language.

1 Introduction

In addition to tangible cultural heritage, intangible cultural heritage is becoming increasingly important, for safeguarding, and protecting traditions and knowledge that each of us identifies as part of our own culture and being as individuals. UNESCO (UNESCO, 2003), in its lists of intangible heritage to protect and safeguard, has several elements related to food, such as the *Art of Neapolitan 'Pizzaiuolo'* or *the Mediterranean diet.* Food products, diet, processing, recipes are an integral part of the cultural identity of people and communities. In particular, traditional recipes are among the elements handed down from one generation to the next and offer a strong link with the territory. They usually are stored in archives, to better preserve and safeguard them.
To exploit the full potential of these archives, easy access modes have been studied to offer all web users, whether they are simply curious or passionate scholars of the subject, tools that provide a complete view of the domain(s) covered. The presence of lists of terms, adapted to the actual archive content, allows us to offer an overview of the topics and to optimize the search and visualization tools. In some specific domains such as photography or architecture, there are glossaries or terminology resources offered by museums or accredited institutions (Research, s.d.). For other domains, as for intangible heritage in general, and recipes in particular, in which these glossaries lack, the only way to get these resources is from scratch or improve existing ones, by experts in the field. It is a time-consuming activity, which requires in-depth knowledge and a systematic mentality.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 10, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2020. ISSN: 1314-4006, eISSN: 2535-0366

In this work, still in progress, we propose a machine-learning approach for automatic identification of domain terms from general and specialized sources. The basic idea of our approach is twofold. First, domain-specific concepts in the documentation are often used as identifiers in specific structural relations in the sentences, and more frequent in specialized documents than in general ones.

Second, integrating different methods, that identify and characterize contents, to exploit the semantic and syntactic aspects of the documents and the contexts in which the terms are used, allows us to identify domain specific terms and therefore to create a more complete thesaurus.

The paper presents a novel method to address the problem of creating flat specific domain thesaurus from texts in the Italian language. One of the strengths of the method is its flexibility: by integrating machine learning methods, with statistical methods, and using pre-trained word embedding and resources easily available on the web, it is possible to (semi-)automatically create glossaries in other domains.

The paper is structured as follows: after the presentation of related works in Section 2, our approach is described in its logical component blocks in Section 3; some experimentation and implementation specifications and choices are detailed in Section 4. Then some preliminary results, with a brief discussion, are presented.

2 Related Works

Automatic extraction of terms has been studied in domains as Information retrieval and text mining, making use of natural language tools and methods. (Artese & Gagliardi, 2014) integrates WordNet to create multilingual glossaries. (Wang, 2019) proposes a learning-based approach for the automatic construction of domain glossary. (Arora, Sabetzadeh, & Briand, 2016) presents a tool-supported approach for extracting candidate glossary terms from natural language requirements and grouping these terms into clusters based on relatedness. Automated glossary construction is having great importance as an aid in navigating and browsing not only archives but also for the textbook. To extract glossary-worthy terms, different strategies have been tested and adopted: machine learning algorithms (Kulkarni, 2018), different architectures of featureless deep learning approaches, including both supervised and semi-supervised models (Khosla, 2017) or deep learning approach (Singh, 2019).

3 Our Approach

The main feature of this approach is the integration of different methods: machine learning classification methods working on the semantic representation of candidate terms, word embeddings models, such as Word2Vec or GloVe able to capture the semantics of words and their context, and a computation of the degree of specialization of a term (called generality).

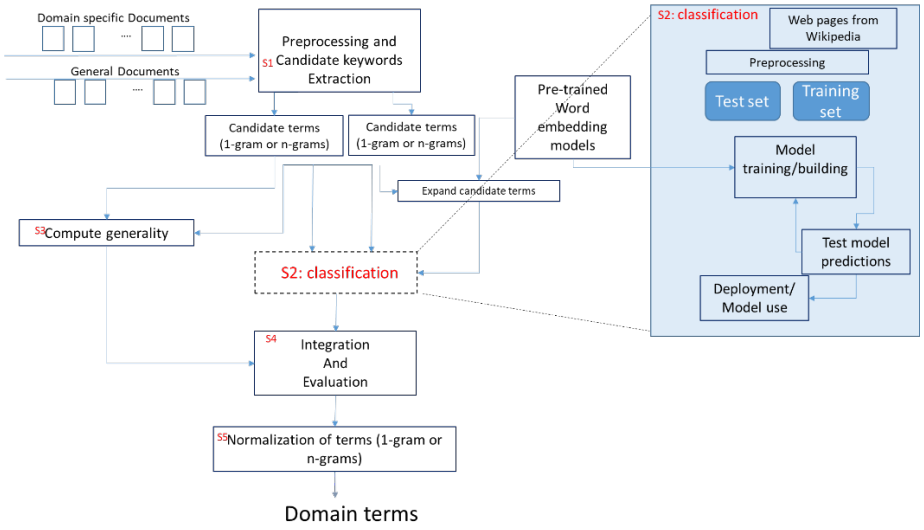

Fig. 1. Approach overview.

An overview of our approach is presented in figure 1: it is composed of six phases, more detailed below.

S1: The Preprocessing Phase aims to eliminate or limit useless or noisy information: the text is tokenized and annotated (pos tagged) to produce candidate terms. The tool that performs best for the Italian language is TreeTagger (Schmid, 2013).

S2: Classify Candidate Terms, as belonging /not belonging to a specific domain. Classification methods underlie a similarity/distance measure to evaluate the belonging of observations to a category from a set of categories defined ahead, based on the training set of data whose category is known.

The method used here to calculate relatedness among words/sentences is based on word embeddings. Word embedding technique is one of the most popular ways to represent document terms, in which words with similar meanings have a similar representation, being able to capture the context of a word in a document, the semantic and syntactic similarity, the relation with other words, etc.

Each word is represented as a real value vector in a predefined vector space, and the vector values are learned in a way that resembles a neural network, so the technique is often inserted in the field of deep learning.

Different models have been developed since 2013 when the first models appeared: we use here word2vec and GloVe.Word2Vec is one of the most used techniques to learn word embedding using shallow neural networks[1] (Mikolov, 2013). The Global

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., & Zweig, G. (2013). word2vec. URL https://code. google. com/p/word2vec

Vectors for Word Representation, or GloVe (Pennington, 2014), algorithm is an extension to the Word2Vec method for efficiently learning word vectors.

Expand candidate terms: as our aim here is to define a set of domain terms, we are interested in enlarging as much as possible the candidate terms to be classified as belonging to the domain of interest. For each original candidate term, the n most similar terms, according to the word embedding model adopted, were classified too.

S3: Compute Generality. To ensure the quality of the extracted terms, we further refine the results of candidate term extraction, eliminating general terms. We use the following equation to estimate the generality of a term,

$$Generality(t) = \frac{freq_g(t)}{freq_d(t)} = \frac{\frac{Occ_g(t)}{TOcc_g}}{\frac{Occ_d(t)}{TOcc_d}} \quad (1)$$

where freq(t) is the frequency of term t, Occ(t) is the occurrence number and TOcc is the terms number-These values are computed for general domain (g) and specific domain (d).

S4: Integration and evaluation. Candidate terms, expanded candidate terms, and generality computation of terms can give different membership belonging values: a phase able to take into account these values to produce a judgment is there required.

S5: normalization of terms: a post-processing phase, to lemmatize and/or to stem extracted terms is required, if in the pre-processing phase candidate terms have been left "as is" in the input streaming.

The result of this method is a list of domain-specific terms, in single or compound form. The grammatical forms of the terms should be restricted to nouns and verbs, in case of 1-gram, and nouns and adjectives, and verbs for n-grams.

4 Experiments

4.1 Datasets, Domain Specific Documents and Websites

Our primary goal is to create a list of terms related to food, in a broad vision, including ingredients, tools, actions, and tradition-related terms, in the Italian language. Different sets of data have been used to implement our approach.
For training and testing the classifiers, we scraped web pages from Wikipedia, starting from the 9 root categories, plus the food-related category. Starting from each category, the scraper tool extracted all the pages of that category, in a recursive way, for a depth of k level (for this experiment, k has been set to 2, a balanced trade-off between the total number of documents to be processed and the variety of terms to consider).

To evaluate the results of the classifiers, it has been used a human-created list of terms, containing ingredients, tools, and actions, taken from the web and manually integrated with missing terms.

The generality value is computed as a ratio between the use of a term in a specific domain and its general use. To find the food-related terms and their occurrences, in a "recall-oriented" way, we extracted terms (1-gram and n-grams/ nouns, nouns+adjectives and verbs), from free available recipes e-books. The list of general terms and their occurrences have been taken from the web[2].

CookIT[3], a web portal aimed at collecting, and sharing Italian traditional recipes related to regional cuisine, has been used as a further source of food-related terms (Artese & Ciocca, 2019).

With the exclusion of the lists of terms, the other datasets have been pre-processed, to extract candidate terms. After some tests, terms have been pos-tagged, with TreeTagger software, to extract 1-grams / n-grams.

4.2 Classifiers and Generality Computation

Two standard classification algorithms have been proposed and tested: Logistic Regression and K-nearest neighbors.

Using the Wikipedia dataset, split into training data and test data, we trained models of the two classifiers chosen for this experiment, testing different parameter values to obtain the best results. The models have then been validated both on the test set and on the food-related lists, defined above.

The classifiers use the word embedding vectors for both pages and titles of the Wikipedia dataset, belonging to the chosen categories: each page/title computed as the mean of all terms. In this way, distance measures among all the documents can be computed to classify terms and documents.

Starting from some 5000 Wikipedia pages for each root category + specific domain, several classification models have been tested: the more performing ones are those that classify in two categories, using as training set items belonging to the domain-specific category and another root category, e.g. philosophy or mathematics. Other tests using all the root categories have given poor results, probably due to the unbalancing of the target category, respect the others, or to the subtler classification.

The use of the classifiers on the CookIT dataset shows that Knn with k=2 is the method that extracts the higher number of terms as food-related, while logistic regression the lower. In detail, candidate terms are 1819, and Knn2 identifies 1597 as food-related, knn5 1512, and logistic regression 1452. Using the expansion, the number of terms explode: as an example for knn2, the expanded number of terms may become, in some cases, more than 10 thousand. The use of the generality helps in strengthening membership belonging to the domain. In fact, eliminating those terms whose generality is higher than a threshold (6 in our implementation - value chosen based on trial term

[2] https://github.com/hermitdave/FrequencyWords/tree/master/content/2018

[3] Cookit portal: http://arm.mi.imati.cnr.it/cookIT

refinement using different thresholds), the resulting terms are 973, most of which in common the classified terms.

Using general candidate terms, not extracted from food-related sources, knn2 (and partially knn5) mistakenly classifies in the thesaurus terms not belonging to it. Logistic Regression classifier performs best in this case. In this case, generality helps in eliminating those terms erroneously inserted in the list.

The results discussed uses the pre-trained w2v embedding model for Italian.

5 Conclusions

The paper presents a machine learning approach, work in progress, for the automatic identification of domain terms from general and specialized sources. The method is easy to implement and integrates standard algorithms, pre trained word embedding models, and web scraped documents to train the classification algorithms and compute generality. Preliminary tests have produced promising results: we will work on fine tuning of the classifiers, the datasets for their training and the generality computation.

References

Arora, C., Sabetzadeh, M., & Briand, L. &. (2016). Automated extraction and clustering of requirements glossary terms. *IEEE Transactions on Software Engineering*, 918–945.

Artese, M. T., & Ciocca, G. &. (2019). CookIT: A Web Portal for the Preservation and Dissemination of Traditional Italian Recipes. *International Journal of Humanities and Social Sciences*, 171–176.

Artese, M., & Gagliardi, I. (2014). Multilingual Specialist Glossaries in a Framework for Intan-gible Cultural Heritage. *International journal of heritage in the digital era (Online)*. doi:https://journals.sagepub.com/doi/abs/10.1260/2047-4970.3.4.657

Khosla, K. J. (2017). *Featureless Deep Learning Methods for Auto-mated Key-Term Extraction.*

Kulkarni, A. &. (2018). *Automated glossary construction of a biology textbook.* Stanford CS229: Machine Learning, Fall.

Mikolov, T. S. (2013). Distributed represen-tations of words and phrases and their compositionality. *Advances in Neural Information Pro-cessing Systems*, 3111–3119.

Pennington, J. S. (2014). Glove: Global vectors for word repre-sentation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (p. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Research, G. (s.d.). Tratto da Art & Architecture Thesaurus: from https://www.getty.edu/research/tools/vocabularies/aat/

Schmid, H. (2013). Probabilistic part-of speech tagging using decision trees. *New methods in language processing*, 154.

Singh, M. &. (2019). *Automatic Extraction of Textbook Glossaries Using Deep Learning.*

UNESCO. (2003). *Intangible Heritage Home*. Tratto da https://ich.unesco.org/en/home

Wang, C. P. (2019). A learning-based approach for automatic construction of domain glossary from source code and documentation. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (p. 97-108). Tallinn, Estonia: Association for Computing Machinery.

Received: June 19, 2020
Reviewed: July 12, 2020
Finally Accepted: July 30, 2020