# Preliminary Concepts for Requirements Mining and Classification using Hidden Markov Model

**László Tóth**

**Abstract:** Requirements specifications are crucial documents of the software development. These documents are created based on the expectations of stakeholders and the requirements of various regulations. The expectations of stakeholders might contain some ambiguities, inconsistencies and also some contradictions especially comparing with regulations. Creating specifications which do not contain issues mentioned above is one of the most important tasks of the business analysts. Reconciling the expectations and requirements can be a demanding task particularly in case of a complex system. Several attempts have been made to support the duties of business analysts using computer-aided natural language processing methods for requirements engineering. One of the most important steps during the elicitation process is the classification of the requirements collected from stakeholders considering that these requirements will be part of the formal and specific models. Investigations devoted to this task have achieved remarkable results using supervised and semi-supervised methods. However, these models need somehow prepared requirements in order to use them as their inputs. The approach presented in this article focuses on extracting and classifying requirements from unstructured documents. Hidden Markov Models are utilized in various field of natural language processing and their usability already has been proven. The idea of using HMM for processing requirements is stemmed from the success of using it for those tasks where extracting information often hidden in unstructured texts is a crucial part of the particular task. Using HMMs, which is a novel approach to processing texts containing requirements, can help also utilize various linguistic features of the sentences that could be obtained with difficulty by classical processes.

**Keywords:** HMM, HHMM, AHMM, NLP, Requirements, Ontologies

## Introduction

Requirements engineering is a vital part of the software development process. This task is accomplished by business analysts based on regulations of the specific business area and interviews which are come from stakeholders. The quality of the resulting specification has a significant impact on the software quality and also the effort needed to develop the software product. Specifications written poorly have heavy consequences for the quality, the budget and also can make a hinder for the maintenance [Firesmith, 2007].

The basis of the requirements specification is made up documents and memos written in natural languages, therefore natural language processing methods can play an important part for supporting business analysts in elicitation processes. Several investigations with considerable results have been made in order to extract and classify requirements from these documents such as [Rashwan et al., 2013], [Casamayor A, 2010], [Cleland-Huang et al., 2007], [Robeer et al., 2016] and [Abad et al., 2017]. Researchers have been focused on the categorization of requirements using supervised and semi-supervised methods. In order to support the finding of the specific texts in unstructured documents, some researchers have applied ontology-based classification methods focusing on the recognition of non-functional requirements [Al Balushi et al., 2007], [Rashwan et al., 2013]. Rashwan et al. [Rashwan et al., 2013] have also created an annotated corpus which is based on the ISO/IEC 9126-1:2001 standard to support classification processes for requirements when there are only a few labeled examples available. [1]

Ontology-based methods might have an important role in the further researches to find an adequate method for minimizing the manual processes regarding requirement elicitation. Ontologies make an important part also in our idea which is based on utilizing the power of Hidden Markov Models in extracting and classifying requirements originated from various textual inputs. There has been no known research focusing on utilizing Hidden Markov Model for requirements engineering tasks, therefore our research as using it for this purpose is a novel approach for harnessing the power of the model.

[1] This standard has been revised and a new standard was published in 2011 which is the ISO/IEC 25010:2011.

## Concepts

As mentioned in the previous section several experiments have been accomplished in order to utilize various a priori information during the process of requirement elicitation. Rashwan et al. have built up an ontology-based corpus which content based on the ISO 9126 standard. The content of the corpus was annotated manually and those examples which annotation was agreed among the participants have been retained in the database called by authors as gold standard [Rashwan et al., 2013]. Databases like gold standard mentioned previously can support elicitation process in different business environments without using manual labeling.

In regards to lack of the labeled examples semi-supervised learning approach can play a crucial role in requirements classification and also their identification. Despite the supervised learning methods have decreased the manual work significantly, tools based on these processes have not spread yet. Semi-supervised learning processes and tools are extremely useful where there are only a few labeled examples available. Casamayor et al. have applied Expectation Maximization strategy along with Naive Bayes method for classification of non-functional requirements and their result has surpassed the accuracy and the precision of supervised methods despite the reduced number of labeled examples [Casamayor A, 2010]. Their experiments have shown that increasing the amount of the labeled examples the accuracy and the precision of the model is increasing as well. Semi-supervised methods by their performance and usability can be good candidates for tools supporting the elicitation process however their results have to be checked manually.

Combining semi-supervised methods and ontology-based approach the increase of performance and usability is expected. Furthermore considering that requirements are given as a textual information which is a sequence of linguistic elements and processing it using methods designed for sequence processing we might extract information which is not obtainable for classical methods. This approach provides some other possibilities such as using n-gram samples or exploiting the relationship among the parts of speech or considering also the modality of the given sentence. Hidden Markov Models are often used in various field of sequence processing such as speech recognition, part of speech tagging, named-entity recognition, text summarization, text classification or topic segmentation [Gao and Zhu, 2013], [Al-Anzi, 2017]. The aim of our research is to examine the possibility of using Hidden Markov Models and ontology-based databases for requirements extraction and classification processes.

## HMM for Requirements Classification

Requirements classification can be considered as one of the areas of utilizing NLP techniques for requirement engineering. Using HMM for requirement engineering is a novel approach, where the classification is that specific field where this attempt is to be applied first for the sake of comparison with results achieved by using classical methods.

Hidden Markov Models are statistical models where the system which is modeled by HMM can be described as a sequence of observed and unobserved states. In these models, the unobserved states of the given sequence have a property called Markov property which means that the given state of the process only depends on the previous state. This property can be written formally as:

$$P(X_m = x_m | X_{m-1} = x_{m-1}, ..., X_1 = x_1) = P(X_m = x_m | X_{m-1} = x_{m-1})$$

Observable sequences are the series of words which is the sentence to be classified. Defining hidden states for our purpose is considered as a research question. The simplest answer is that we can use classes of requirements for this purpose. We can construct binary classifiers for each class where hidden states can be reduced to only two states. We can also utilize the hierarchical hidden Markov models (HHMM) in order to process the results given by binary classifiers and making the final decision. We can construct models with much more hidden states in order to exploit other linguistic features of the given sentence such as modality or the relationship among the parts of speech.

In the Markov models, two kinds of probabilities are defined. The first one is the transition probability which is conceived between two hidden states, and the emission probability which is the probability of the observed element given in a specific state. The emission probability can be calculated based on ontologies or labeled examples. What other information can be used from ontologies in order to construct our model is also an open question.

Another interesting approach to classify requirements is using Aspect Hidden Markov Models (AHMM) which was proposed by David M. Blei and Pedro J. Monero for topic segmentation tasks in their *Topic Segmentation with an Aspect Hidden Markov Model* article [Blei and Moreno, 2001]. Their work is based on Hoffman's aspect model [Hofmann, 1999]. The mentioned model can be described as a group of probability distributions over two discrete random variables. These random variables are the labels of the specific requirements and the words of the given sentence in our case. The supposition of the model is that the sentence and words are independent of each other, given a specific label. Then the joint probability can be written as [Blei and Moreno, 2001]:

$$P(d, w, z) = P(d|z)P(w|z)P(z)$$

where d denotes the document or the sentence in our case, w denotes the words and z denotes the labels. Although the given supposition is strong this model has performed well in the topic segmentation tasks [Blei and Moreno, 2001].

The various HMMs demand that their inputs be preprocessed according to the attributes of the particular model and the information to be extracted. To find appropriate methods for preprocessing texts containing requirements without losing important information is another research question. Procedures applicable for topic segmentation might be a good starting point.

Due to the lack of relevant researches corresponding to apply Hidden Markov Models for requirements engineering tasks defining a model for requirements classification is an open topic. HMMs can be used for supervised learning if there are a lot of labeled examples available which is unfortunately not the case. Using ontologies, semi-supervised learning can be a reasonable choice. We can construct binary classifiers for each class and these classifiers can be used for a hierarchical model which is applicable for making the final decision. We can also consider those cases where sentences can be classified into more than one classes.

# Conclusion

Using Hidden Markov Model for requirements classification is a novel approach to this problem. There are plenty of open questions emerging at the beginning of the research. The first question is the architecture of the model. We have presented some ideas in this article regarding this question however a lot of other constructions can be taken into. One of our future goals is to find an efficient architecture which can support both the extraction and the classification processes.

The second problem is the stability of the model. Computed probabilities regarding using sparse matrices can be very small therefore smoothing techniques are to be applied. Choosing the appropriate smoothing method which fits best for our purpose is another important research question.

We propose further possibilities to improve usability and performance of the extraction and the classification processes such as exploiting relationships among the part of speech, using n-gram models and harnessing the information provided by ontologies. In the long term, the main purpose of our research is developing a tool which can be used to support the duty of business analysts in requirements elicitation processes.

### References


[Abad et al., 2017] Abad, Z. S. H., Karras, O., Ghazi, P., Glinz, M., Ruhe, G., and Schneider, K. (2017). What Works Better? A Study of Classifying Requirements. In *Proceedings - 2017 IEEE 25th International Requirements Engineering Conference, RE 2017*, pages 496–501.

[Al-Anzi, 2017] Al-Anzi, F. S. (2017). STATISTICAL MARKOVIAN DATA MODELING FOR NATURAL LANGUAGE PROCESSING. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 7(1).

[Al Balushi et al., 2007] Al Balushi, T. H., Sampaio, P. R. F., Dabhi, D., and Loucopoulos, P. (2007). ElicitO: A Quality Ontology-Guided NFR Elicitation Tool. In *Requirements Engineering: Foundation for Software Quality*, pages 306–319. Springer Berlin Heidelberg.

[Blei and Moreno, 2001] Blei, D. M. and Moreno, P. J. (2001). Topic segmentation with an aspect hidden Markov model. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 01*, 12(2-3):343–348.

[Casamayor A, 2010] Casamayor A, Godoy D, C. M. (2010). Identification of non-functional requirements in textual specifications: A semi-supervised learning approach. *Information and Software Technology*, 52(4):436–445.

[Cleland-Huang et al., 2007] Cleland-Huang, J., Settimi, R., Zou, X., and Solc, P. (2007). Automated classification of non-functional requirements. *Requirements Engineering*, 12(2):103–120.

[Firesmith, 2007] Firesmith, D. (2007). Common requirements problems, their negative consequences, and the industry best practices to help solve them. *Journal of Object Technology*, 6(1):17–33.

[Gao and Zhu, 2013] Gao, X. and Zhu, N. (2013). Hidden Markov model and its application in natural language processing. *Information Technology Journal*, 12(17):4256–4261.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 50:2.

[Rashwan et al., 2013] Rashwan, A., Ormandjieva, O., and Witte, R. (2013). Ontology-Based Classification of Non-functional Requirements in Software Specifications: A New Corpus and SVM-Based Classifier. In *2013 IEEE 37th Annual Computer Software and Applications Conference*, pages 381–386. IEEE.

[Robeer et al., 2016] Robeer, M., Lucassen, G., van der Werf, J. M. E. M., Dalpiaz, F., and Brinkkemper, S. (2016). Automated Extraction of Conceptual Models from User Stories via NLP. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pages 196–205. IEEE.