USING NATURAL LANGUAGE PROCESSING FOR AUTOMATIC EXTRACTION OF ONTOLOGY INSTANCES

Carla Faria
Federal Institute for Education, Science, Technology of Maranhão, Computer Science Department
São Luiz, Maranhão, Brazil

Rosario Girardi, Ivo Serra, Maria Macedo, Djefferson Maranhão
Federal University of Maranhão, Computer Science Department, São Luiz, Brazil

Keywords: Ontology, Ontology population, Natural language processing, Knowledge acquisition.

Abstract: Ontologies are used by modern knowledge-based systems to represent and share knowledge about an application domain. Ontology population looks for identifying instances of concepts and relationships of an ontology. Manual population by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and semi-automatic approaches are needed. This article proposes an initial approach for automatic ontology population from textual sources that use natural language processing and machine learning techniques. Some experiments using a family law corpus were conducted in order to evaluate it. Initial results are promising and indicate that our approach can extract instances with good effectiveness.

1 INTRODUCTION

Ontologies are an approach for knowledge representation capable of expressing a set of entities, their relationships, constraints and rules (conditional statements) of a given area (Guarino, Masolo and Vetere, 1999) (Nierenburg and Raskin, 2004). They are used by modern knowledge-based systems for representing and sharing knowledge about an application domain. These knowledge representation structures allow the semantic processing of information and, through more precise interpretation of data, systems have greater effectiveness and usability.

Ontology population is the term used to designate the techniques for extracting and classifying instances of concepts and relationships of an ontology. Manual population of ontologies by domain experts and knowledge engineers is an expensive and time-consuming task. Therefore, automatic or semi-automatic approaches are required.

This paper proposes an approach for automatic ontology population from textual resources based on natural language processing (Allen, 1995) (Dale, Moisl, Somers, 2000) and machine learning techniques (Russel and Norvig, 1995) (Bichop, 2006). It details the process techniques used in corpus creation and identification of candidate instances phases. An experiment conducted to evaluate these techniques is also described. The experiment consists of extracting instances from a corpus of jurisprudence to populate FAMILYLAW, an ontology developed for the family law domain.

This article is organized as follows. Section 2 introduces the ontology definition used in this work. Section 3 summarizes related work. Section 4 gives an overview of the proposed process for automatic ontology population. Section 5 details the identification of the candidate instances phase. Section 6 describes an experiment conducted to evaluate the proposed technique for the identification of candidate instances and finally, section 7 concludes the work.

2 AN ONTOLOGY DEFINITION

Ontologies are formal specifications of concepts in a domain of interest. Their classes, relationships,

Faria C., Girardi R., Serra I., Macedo M. and Maranhão D. (2010).
USING NATURAL LANGUAGE PROCESSING FOR AUTOMATIC EXTRACTION OF ONTOLOGY INSTANCES.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 278-283
DOI: 10.5220/0002904402780283

constraints and axioms define a common vocabulary to share knowledge (Gruber, 1995).
Formally, an ontology can be defined as:

$$O = (C, H, I, R, P, A) \qquad (1)$$

where

$C = C^C \cup C^I$ is the set of entities of the ontology. They are designated by one or more terms in natural language. The set C^C consists of classes, i.e., concepts that represent entities that describe a set of objects (for example, "Person" $\in C^C$) while the set C^I is constituted by instances, (for example "Erik" $\in C^I$).

$H = \{\text{kind\_of}(c_1,c_2) \mid c_1 \in C^C, c_2 \in C^C\}$ is the set of taxonomic relationships between concepts, which define a concept hierarchy and are denoted by "kind_of(c_1,c_2)", meaning that c_1 is a subclass of c_2, for instance, "kind_of(Lawyer,Person)".

$I = \{\text{is\_a}(c_1,c_2) \mid c_1 \in C^I, c_2 \in C^C\}$ is the set of relationships between classes and instances of an ontology, for example, "is_a (Anne,Client)".

$R = \{\text{rel}_k(c_1,c_2,\ldots, c_n) \mid \forall i, c_i \in C\}$ is the set of ontology relationships that are neither "kind_of" nor "is_a". Some examples are "represents(Lawyer, Client)" and "represents(Erik, Anne)".

$P = \{\text{prop}^C(c_k,\text{datatype}) \mid c_k \in C^C\} \cup \{\text{prop}^I(c_k,\text{value}) \mid c_k \in C^I\}$ is the set of properties of ontology entities. The relationship prop^C defines the basic datatype of a class property while the relationship prop^I defines its instance value. For instance, subject (Case, String) is an example of a prop^C property and subject (Case12, adoption) is an example of a prop^I property.

$A = \{\text{condition}_x \Rightarrow \text{conclusion}_y(c_1,c_2,\ldots, c_n) \mid \forall j, c_j \in C^C\}$ is a set of axioms, rules that allow checking the consistency of an ontology and infer new knowledge through some inference mechanism. The term condition$_x$ is given by $\text{condition}_x = \{(\text{cond}_1,\text{cond}_2,\ldots,\text{cond}_n) \mid \forall z, \text{cond}_z \in H \cup I \cup R\}$. For example, "applied_to(Defense_Argument22, Case12), similar_to(Case12, Case13) ⇒ applied_to (Defense_Argument22, Case13)" is a rule that indicates that if two legal cases are similar then, the defense argument used in one case could be applied to the other one.

As an example, consider a very simple ontology describing the domain of a law firm (Figure 1), which has lawyers responsible for the cases of clients they serve.

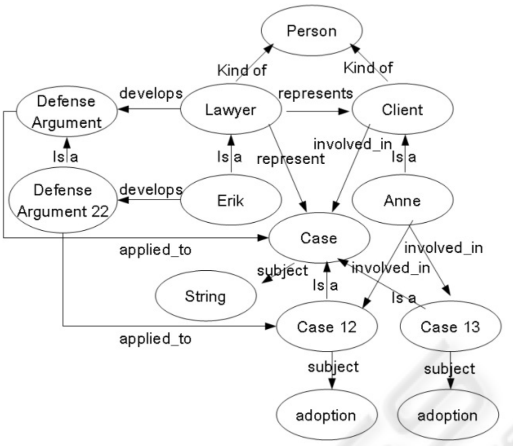


Figure 1: Example of an ontology of a law firm.

According to the previous ontology definition, from the ontology in the Figure 1, the following sets can be identified.

C^C = {person, lawyer, client, case}.

C^I = {Erik, Anne, Case12, Case13, DefenseArgument22}.

H = {kind_of(Person, Lawyer), kind_of(Person, Client)}.

I = {is_a(Erik, Lawyer), is_a(Anne, Client), is_a(DefenseArgument22, DefenseArgument), is_a(Case12, Case), is_a(Case13, Case)}.

R = {represents(Lawyer, Client), represent (Lawyer, Case), applied_to(DefenseArgument, Case), develops (Lawyer, Defense_Argument), involved_in(Client, Case)}.

P = {subject(Case, String), subject(Case12, adoption), subject(Case13, adoption)}.

A = applied_to(Defense_Argument22, Case12), similar_to(Case12, Case13) ⇒ applied_to (Defense_Argument22, Case13).

3 RELATED WORK

There are two main paradigms in ontology population (Tanev and Magnini, 2006). In the first one, the population of ontologies can be performed using patterns (Hearst, 1998) or analyzing the structure of terms (Velardi, Navigli, Cuchiarelli and Neri, 2005). For example, in the phrase "Umberto Eco is an author", "Umberto Eco" can be considered an instance of the "author" class and therefore there is an "is-a" relationship between these two terms. In the second paradigm, the task is addressed using contextual features (Cimiano and Volker, 2005).

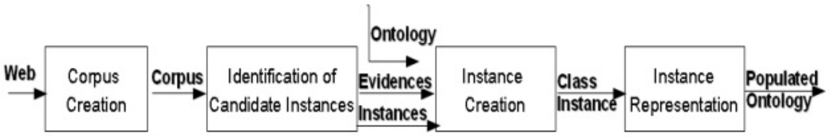


Figure 2: An automatic ontology population process.

Context feature approaches use a corpus to extract features from the context in which a semantic class tends to appear (Fleischman and Hovy, 2002). For example, to classify John Smith as a teacher, the features that indicate the classification of "John Smith" as a teacher could be the verb "to teach".
A hybrid approach using both pattern-based, term structure, and contextual feature methods is presented in (Cimiano, Pivk, Thieme and Staab, 2005).

4 AN OVERVIEW OF A PROCESS FOR AUTOMATIC ONTOLOGY POPULATION

The process proposed in this paper is a hybrid approach, considering that it uses both contextual features and recognition based on lexical patterns. The tasks of the process are supported by supervised machine learning and natural language processing techniques. It consists of four phases: Corpus Creation, Identification of Candidate Instances, Instance Creation and Instance Representation (Figure 2).

The corpus creation phase aims at capturing documents through a web crawler to construct a corpus to be used in the following phase.

The identification of candidate instances phase looks for structuring the documents in the corpus by applying techniques of natural language processing and statistic measures in order to extract instance candidates.

The instance creation phase aims at classifying an instance into a particular class by applying machine learning techniques. The products of this phase are: the CI set of the ontology definition in section 2 and the set of all "is_a" relationships between classes and instances, corresponding to the I set of the ontology definition in section 2. This phase has as input: evidences, candidate instances and an ontology. It is divided into three tasks, training set construction, building a classifier and classification, as shown in Figure 3.

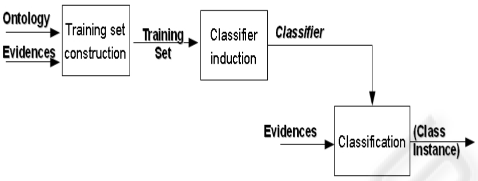


Figure 3: The instance creation subprocess.

The training set construction task generates a set of positive examples of the type (evidence, class), where "evidence" is a syntactic dependence obtained in the identification of candidate instances phase and the class name is obtained from the ontology to be populated; "class" is the class of the domain ontology where that candidate instance should be classified. The instances are assigned to their respective classes of the ontology, this task is performed manually.

The classifier induction task submits the training set to a machine learning algorithm, that induces a classifier matching the best approximation of the "is_a" relationship. The classifier used was Naïve-Bayes (Mitchell, 1997.), the simplest of Bayesian classifier (Witten and Frank, 2005).

Instances are associated with ontology classes, generating as a result a set of pairs (instance, class) in the classification task.

The instance representation phase aims at representing instances in a particular ontology specification language.

5 THE IDENTIFICATION OF CANDIDATE INSTANCES PHASE

The identification of candidate instances phase consists of the tasks illustrated in Figure 4. Each one of the tasks is following described with an example using the simple corpus of Figure 5.

The tokenization/normalization task aims at dividing the text into tokens (Figure 6) and identifying standards such as dates, times, hours, among others. For instance, the tokens "five" and "two" in the corpus of Figure 5 are of the standard time.

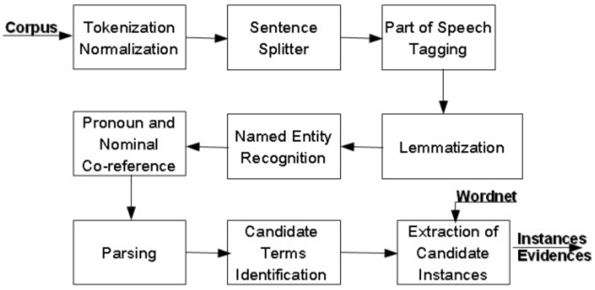
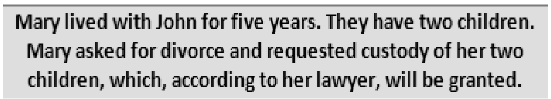
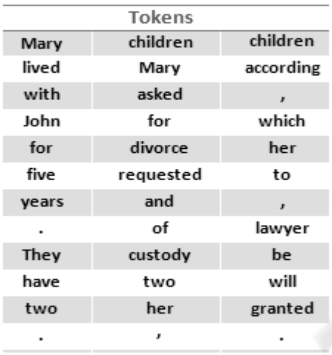

Corpus
Tokenization Normalization
Sentence Splitter
Part of Speech Tagging
Lemmatization
Named Entity Recognition
Pronoun and Nominal Co-reference
Parsing
Candidate Terms Identification
Wordnet
Extraction of Candidate Instances
Instances
Evidences

Figure 4: Identification of candidate instances phase.



Mary lived with John for five years. They have two children. Mary asked for divorce and requested custody of her two children, which, according to her lawyer, will be granted.

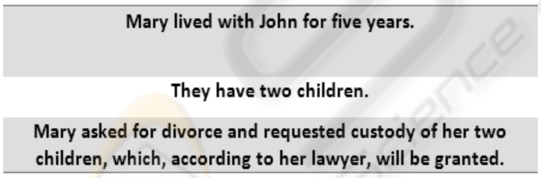
Figure 5: A simples corpus in the family law domain.



Tokens		
Mary	children	children
lived	Mary	according
with	asked	,
John	for	which
for	divorce	her
five	requested	to
years	and	,
.	of	lawyer
They	custody	be
have	two	will
two	her	granted
.	,	.

Figure 6: Tokens extracted from the corpus of Figure 5.

The sentence splitter task aims at identifying sentences in the text, as shown in Figure 7.



Mary lived with John for five years.

They have two children.

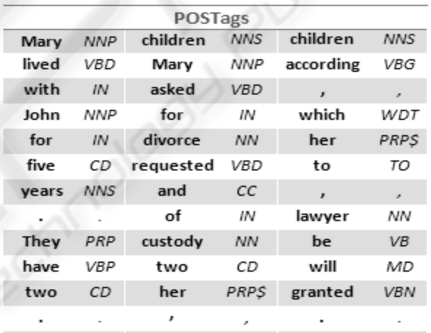
Mary asked for divorce and requested custody of her two children, which, according to her lawyer, will be granted.

Figure 7: Sentence splitter from the corpus of Figure 5.

The Part of Speech Tagging (POS tagging) task looks for assigning syntactic categories to each token, according to the Penn TreeBank set of tags (Marcus, Santorini, Marcinkiewicz 1993), as shown in Figure 8.

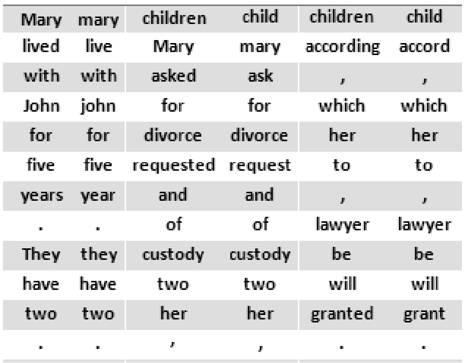
In the lemmatization task tokens are reduced to their basic forms. For example, nouns are represented in their male single form and verbs in the infinitive form, as shown in Figure 9.

The Named Entity Recognition task identifies names that refer to unique objects in the world such as names of person, organizations and places. For instance, in the corpus of Figure 5 the only two named entities are "Mary" and "John".



POSTags					
Mary	*NNP*	children	*NNS*	children	*NNS*
lived	*VBD*	Mary	*NNP*	according	*VBG*
with	*IN*	asked	*VBD*	,	,
John	*NNP*	for	*IN*	which	*WDT*
for	*IN*	divorce	*NN*	her	*PRP$*
five	*CD*	requested	*VBD*	to	*TO*
years	*NNS*	and	*CC*	,	,
.	.	of	*IN*	lawyer	*NN*
They	*PRP*	custody	*NN*	be	*VB*
have	*VBP*	two	*CD*	will	*MD*
two	*CD*	her	*PRP$*	granted	*VBN*
.	.	,	,	.	.

Figure 8: Pos tagging from the corpus of Figure 5.

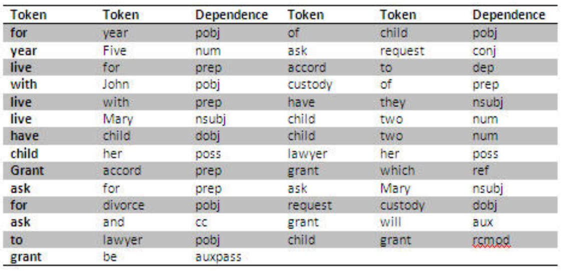


Mary	mary	children	child	children	child
lived	live	Mary	mary	according	accord
with	with	asked	ask	,	,
John	john	for	for	which	which
for	for	divorce	divorce	her	her
five	five	requested	request	to	to
years	year	and	and	,	,
.	.	of	of	lawyer	lawyer
They	they	custody	custody	be	be
have	have	two	two	will	will
two	two	her	her	granted	grant
.	.	,	,	.	.

Figure 9: Lemmatization from the corpus of Figure 5.

The co-reference task identifies both pronoun and nominal co-references. The first one consists of pronouns that refer to previously described entities. For instance, "her" is the pronoun that refers to "Mary" in the corpus of Figure 5. The nominal co-reference consists of nouns that refer to the same entity. For instance, “Barak Obama”, “Mr. Barak Obama” and “President Obama” refer to Barak Hussein Obama, president of the United States.

The parsing task aims at building a parse tree of each sentence in the text and identifying syntactic dependences according to the Stanford dependences among terms (Marneffe and Manning, 2008), as shown in Figure 10.



Token	Token	Dependence	Token	Token	Dependence
for	year	pobj	of	child	pobj
year	Five	num	ask	request	conj
live	for	prep	accord	to	dep
with	John	pobj	custody	of	prep
live	with	prep	have	they	nsubj
live	Mary	nsubj	child	two	num
have	child	dobj	child	two	num
child	her	poss	lawyer	her	poss
Grant	accord	prep	grant	which	ref
ask	for	prep	ask	Mary	nsubj
for	divorce	pobj	request	custody	dobj
ask	and	cc	grant	will	aux
to	lawyer	pobj	child	grant	rcmod
grant	be	auxpass			

Figure 10: Parsing from the corpus of Figure 5.

The candidate terms identification task aims at selecting the terms identified as proper nouns and eliminate others, as proper nouns are probably instances. For instance, “Mary” and “John” could be candidate terms in the corpus of Figure 5.

The extraction of candidate instances task aims at extracting candidate instances based on the Inverse Document Frequency measure (IDF) (Salton and Buckley, 1987) which measures the general importance of a term in the corpus. The normalized inverse document frequency of term i ($IDF_i \in [0, 1]$) is defined as:

$$IDF_i = \log (N / n_i) / \log (N) \qquad (2)$$

where N is the number of documents in the corpus and n_i is the number of documents where the term i occurs.

IDF_i is calculated for each term and those having a value between an experimentally defined interval are selected as candidate instances. Those terms having an IDF_i out of such predefined interval are compared with a set of instances available in the Wordnet lexical database (Felbaum, 1998) are also selected. The final list of candidate instances is composed of the terms selected from the Wordnet and the terms identified in the IDF_i predefined interval. For instance “Mary” and “John” could be candidate instances in the corpus of Figure 5.

6 EVALUATION

An experiment was conducted for an initial evaluation of the effectiveness of the proposed approach for identifying candidate instances already described in section 5.

A corpus composed of 919 documents captured from the site “family.findlaw.com” and containing jurisprudence cases in the domain of family law was used in the experiment performed with the GATE platform (GATE, 2009).

An adaption of the classical measures of recall and precision from the information retrieval area was used for effectiveness evaluation (Dellschaft and Staab, 2006) considering the number of candidate instances that were identified correctly under an interval value of the IDF_i. Precision measures the ratio between the number of relevant instance candidates identified and the number of extracted candidates; and recall, the one between the number of relevant instance candidates identified and the number of instances in the corpus.

$$\text{Precision} = N_{\text{Rel Ident}} / N_{\text{Ident}} \qquad (3)$$

$$\text{Recall} = N_{\text{Rel Ident}} / N_{\text{Rel Corpus}} \qquad (4)$$

where $N_{\text{Rel Ident}}$ is the number of relevant instance candidates identified, N_{Ident} is the number of instance candidates identified, and $N_{\text{Rel Corpus}}$ is the number of instances in the corpus.

Figure 11 show the precision and recall graphics, for different IDF_i values. It can be experimentally observed that, the IDF_i interval [0,56, 0,93] has the best balance between recall and precision. A sample of one thousand terms was analyzed in this interval getting a precision of 54%.

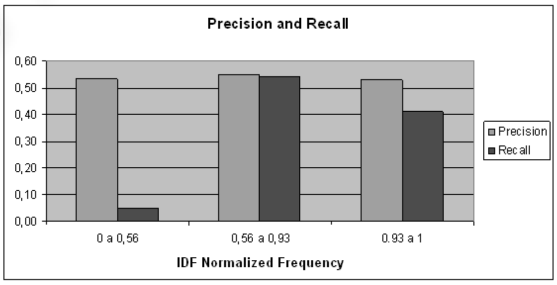


Figure 11: Precision and Recall x IDF_i results.

A second experiment to extract instances from the same corpus using IDF_i and instances taken from the Wordnet out of the selected interval was performed. With a sample of one thousand terms a precision of 64% was obtained which represent an improvement of 10% when compared with the first experiment.

7 CONCLUDING REMARKS

Manual population of ontology by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and semi-automatic approaches are needed.

This article gives an overview of a domain independent process for automatic ontology population from textual resources and details the phase where candidate instances of an ontology are identified. The process is based on natural language processing and supervised machine learning techniques and consists of four phases: corpus creation, identification of candidate instances, instance creation and instance representation.

Two experiments were performed to evaluate the proposed approach. The first experiment used natural language processing techniques and the IDF statistical measure. Candidate instances were extracted from a corpus in the family law domain and a precision of 54% was obtained. The second experiment used the previously described techniques and Wordnet getting an improvement of 10% in the precision value.

The combination of natural language processing techniques and statistical measures seems to be a promising approach for automatic extraction of ontology instances considering the initial results reported here. However, more experimentation is needed.

Currently we are evaluating different supervised machine learning algorithms (Bayesian networks, decision trees and statistical relational learning, among others) in order to select a suitable approach for the classification of instances in ontology classes. We are also evaluating the advantages of combining information extraction techniques with the proposed approach to improve its effectiveness.

ACKNOWLEDGEMENTS

This work is supported by CNPq, CAPES and FAPEMA, institutions of the Brazilian Government for scientific and technologic development.

REFERENCES

Allen, J. 1995. Natural Language Understanding. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.

Bichop, C. M. 2006. Pattern Recognition and Machine Learning, Springer.

Cimiano, P. and Volker, J., 2005. Towards large-scale, open-domain and ontology-based named entity classification. In: *Proceedings of RANLP'05*, p. 166–172, Borovets, Bulgaria.

Cimiano,P., Pivk, A., Thieme, L. S. and Staab, L. S., 2005. Learning Taxonomic Relations from heterogeneous Sources of Evidence. In Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press.

Dale, R., Moisl, H. and Somers, H. L. 2000. Handbook of natural language processing. CRC Press.

Dellschaft, K. and Staab, S. 2006. On how to perform a gold standard based evaluation of ontology learning. In: *Proceedings of the 5th International Semantic Web Conference*, p. 228 – 241, Athens. Springer.

Fellbaum, C., 1998. Wordnet: An Electronic Lexical Database, MIT Press.

Fleischman, M. and Hovy, E., 2002. Fine Grained Classification of Named Entities. In: *Proceedings of COLING*, Taipei, Taiwan.

General Architecture for Text Engineering, 2009, http://gate.ac.uk, December.

Gruber, T. R., 1995. Toward Principles for the Design of Ontologies used for Knowledge Sharing. International Journal of Human-Computer Studies, nº43, pp. 907-928.

Guarino, N., Masolo, C., and Vetere, G. 1999. Ontoseek: Content-based Access to the web. *IEEE Intelligent Systems*, v. 14(3), p. 70-80.

Hearst, M., 1998. Automated Discovery of Word-Net Relations. In WordNet: An Electronic Lexical Database. MIT Press.

Marcus, M., Santorini, B. and Marcinkiewicz, M. 1993. Building a Large Annotaded Corpus of English: Penn TreeBank. Computational linguistics: Special Issue on Using Large Corpora, [S. I.], v. 19, n.2, p. 313 – 330.

Marneffe, M. and Manning, C. 2008. The Stanford typed dependencies representation. In: Workshop on Cross-Framework and Cross-Domain Parser Evaluation, Manchester. *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*. p. 1 - 8.

Mitchell, T. 1997. Machine Learning, Mc Graw Hill.

Nierenburg, S. and Raskin, V. 2004. Ontological Semantics, MIT Press.

Russel, S. and Norvig, P. 1995. Artificial Intelligence: A Modern Approach, Prentice-Hall.

Salton, G. and Buckley, C., 1987. Term Weighting Approaches in Automatic Text Retrieval. Cornell University.

Tanev, H. and Magnini, B., 2006. Weakly Supervised Approaches for Ontology Population. In: *Proceedings of EACL*.

Witten, I. H. and Frank, E. 2005. Data Mining Practical Machine Learning Tools and Techniques, Elsevier 2nd edition.