

Towards an optimal ontology construction

A research project report presented

by

Natalia Ponomareva

to

The Department of Information Systems and Computation

in partial fulfillment of the requirements

for the obtention of

Diploma of Advanced Studies

(Diploma de Estudios Avanzados)

in the subject of

Pattern Recognition and Artificial Intelligence

Polytechnic University of Valencia

Valencia, Spain

January 2008

©2008 - Natalia Ponomareva

All rights reserved.

Author
Natalia Ponomareva
PhD Student

Thesis supervisors
Paolo Rosso
Mikhail Alexandrov

Towards an optimal ontology construction

Abstract

A subject of ontology, its usage, building and evaluation has been becoming one of the most popular research topics for the last ten years. Such a strong interest to this issue can be explained by great expectations that scientists relate to a widespread ontology's usage. Following by Uschold [46] we can emphasize 3 main ontology's purposes:

1. Intelligent communication between different domains.
2. Interoperability among systems and tools.
3. Benefits for knowledge engineering that imply knowledge reuse, reliability, specification, reasoning, new knowledge generation, etc.

The present work aims at focusing on problems of ontology learning. Basically, it tries to solve 2 kinds of problems related to this issue:

1. Problem of term's recognition. Some scientific domains as biological molecular domain are characterized by long descriptive entities. The problem becomes complicated by a quick appearance of new terms and a lack of naming convention. Therefore, simple and well-known methods of term's extraction based on term weighting and terms' cooccurrence do not work in such conditions.

2. Concepts retrieval on a given granularity level. Different tasks require an ontology built for distinct granularity levels. For example, ontology in physics domain used for describing student's education literature needs more general set of concepts than an ontology representing last scientific achievements. Therefore, there is a need to be able to change ontology level of granularity depending on a necessity of a task.

In order to solve the last problem we apply an Inductive method of model self-organization (IMMSO) that is used to search a model of optimal complexity. We elaborate on an artificial example the limits of this method by changing the parameters of initial data. IMMSO is quite a new tool in Computational Linguistic (CL), its previous applications are rather scarce. In our point of view it is a very interesting and useful method and can be adopted on a variety of CL tasks. For instance, in this work we also demonstrate its successful application for dialogue processing.

Summing up the aforesaid we should point out the following contributions of our work:

1. A novel based on Hidden Markov Models (HMM) approach for biomedical Named Entity Recognition (NER) giving good results in case of poor additional information.
2. Comparing performance of different Machine Learning (ML) methods under the same conditions in a biomedical NER task.
3. A formal definition of granularity of domain terms.
4. A method of revealing granularity levels of domain terminology.
5. Elaborating stability of IMMSO for different parameters of initial data.
6. Constructing an empirical formula for estimating client's characteristics in dialogue processing

Contents

Title page	i
Abstract	iii
Table of contents	v
Citations to previously published papers	vii
1 Introduction	1
1.1 Biomedical Named Entity Recognition	1
1.2 Revealing granularity of domain terminology	2
1.3 Overview of the research report	3
2 Ontology learning and evaluation: state of the art	4
2.1 Ontology learning	4
2.2 Ontology evaluation	6
3 Biomedical Named Entity Recognition	11
3.1 Motivation	11
3.2 HMMs and CRFs in sequence labeling tasks	13
3.3 Biomedical NE recognizers description	14
3.3.1 JNLPBA corpus	14
3.3.2 Feature set	16
3.3.3 Two strategies of HMM- and CRF-based models' building	16
3.4 Experiments and discussions	18
3.5 Further work	20
4 IMMSO	21
4.1 The method description	21
4.2 Problem settings	22
4.3 Organization of experiments	23
4.3.1 Models under consideration	23
4.3.2 Artificial data	25
4.3.3 Result evaluation and energetic ratios	25
4.3.4 Methods and criteria	26
4.4 Experiments and results	28

4.4.1	Stability with respect to a data volume	28
4.4.2	Stability with respect to the unexactness of model	29
4.4.3	Stability with respect to the noise	29
4.4.4	Model self-organization for different types of external criteria .	30
4.4.5	Results with the Approximation Technique	30
5	Revealing granularity of domain terminology	33
5.1	Introduction	33
5.2	Corpus-based term granularity	34
5.3	Specificity approximation	35
5.3.1	Entropy-based specificity	36
5.3.2	Standard deviation-based specificity	37
5.4	A method of detecting granularity levels	37
5.5	Experiments and results	39
5.5.1	Corpus characteristics	39
5.5.2	Detecting levels of granularity	39
5.6	Summary and future work	43
6	Constructing empirical models for automatic dialogue processing	44
6.1	Problem setting	44
6.2	Models for parameter estimation	45
6.2.1	Numerical indicators	45
6.2.2	Example	46
6.2.3	Numerical models	47
6.3	Application of IMMSO	47
6.4	Experiments	49
6.5	Conclusions	50
7	Conclusions and future work	51
	Bibliography	53

Citations to previously published papers

Large portions of Chapters 3, 5 and 6 have appeared in the following papers:

Ponomareva N., Blanco X. Example-based empirical formula for politeness estimation in dialog processing. NooJ conference. Barcelona, Spain. 2007

Ponomareva N., Pla F., Molina A., Rosso P. Biomedical Named Entity Recognition: A poor knowledge HMM-based approach. In: Proc. 12th Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2007, Springer-Verlag, LNCS(4593), pp. 382-387, 2007

Alexandrov M., Blanco X., Ponomareva N., Rosso P. Constructing empirical models for automatic dialogue parametrization. In: Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 455-463, 2007

Ponomareva N., Rosso P., Pla F., Molina A. Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task. In: Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, September 27-29, pp. 479-483, 2007

Chapter 1

Introduction

Ontology are considered to be a base of Semantic Web technology because they provide intelligent interaction between applications and agents on the semantic level of communication instead of just lexical or syntactic one. A term “ontology” was originally adopted from philosophy where it means the study of being or existence. Since this notion was introduced into an area of Artificial Intelligence many formal definitions of ontology has appeared. One of the most famous and well-known one was proposed by T. Gruber who defined an ontology as “ an explicit specification of a conceptualization” [17]. We may find two principal parts in this definition characterized a notion of ontology. First of them is “explicit specification”, which highlights that knowledge is represented in an explicit form. “Conceptualization” implies a view of the world or a concrete domain depending on a subject under consideration. In other words, ontologies are explicit, unambiguous, consistent structural representations of some field that may involve concepts, attributes, relationships and constraints.

This work addresses some important issues related to ontology learning. First, we investigate a topic of term recognition in biomedical domain that still remains a challenging task because of the complex properties of biomedical entities. Then, we research a possibility of ontology construction for a given level of granularity.

1.1 Biomedical Named Entity Recognition

The motivation for solving this problem can be explained by the following factors:

1. Biomedical domain has been getting a quick development during the last 10-15 years.
2. A huge amount of new biomedical terms have been appeared.
3. Biomedical terms have quite complex structure.

Some important phenomena of biomedical terms that cause difficulties of their recognition are [49]:

1. Different writing forms existing for one entity.
2. Very long descriptive names.
3. Term share.
4. Cascaded entity problem.
5. Abbreviations.

In this work, we research two ML methods, namely, Hidden Markov Models (HMM) and Conditional Random Fields (CRF) for biomedical Named Entity Recognition (NER). Each of these methods has both benefits and disadvantages although CRFs are conceived as more successful in sequence labeling tasks. One of the main disadvantages of HMMs is that it is impossible to incorporate additional features into the model. In order to solve this problem we propose a technique called state specialization that allows to add some information about words' properties into hidden states. The constructed model outperforms all existing HMM-based classifiers exploited the same amount of a priori information. Moreover, comparison of the obtained results given by both classifiers does not show a strong advantage of the CRFs. CRFs outperform HMMs when comparing F-score while correspondence of recalls demonstrates that HMM-based classifier tends to decrease a second order error, which is preferable, in many cases.

1.2 Revealing granularity of domain terminology

This work is inspired by an idea of ontology construction that could fit a given task. One of the ontology properties depending on a task is an ontology granularity level: logically, different tasks require their own levels of details. Granularity of ontology can be expressed through granularity of ontology concepts and expressiveness of ontology relationships. In our work, we only deal with the granularity of terms without taking into account relationships between them. We define a notion of granularity using a notion of term specificity. In fact, intuitively coarse-grained terms must be popular and well-known terms and, therefore, they cannot be very specific. The opposite is true for fine-grained terms: they must be rare and highly-specialized and, thus, very specific terms.

Our method of revealing granularity levels of terms consists of two steps:

1. Approximation of specificity by one of term-weighting schemes.
2. Application of IMMSO for finding transition points between adjacent granularity levels.

We propose two term-weighting schemes for solving the first problem:

1. An information-theoretical metrics exploiting an entropy of a given term throughout a document collection as a measure of its specificity: the higher entropy a term has the less specific it is.
2. A metrics based on deviation of term distribution inside a document collection. Logically, more specific terms have higher level of fluctuation from their average frequencies whereas the distribution of general terms tends to be more uniform.

We tackle the second problem of discovering boundaries between granularity levels by use of the IMMSO optimization scheme. Briefly, it consists of dividing a data collection into training and control data sets. When using IMMSO we are interested in searching for a global minimum of some external criterion built upon these two data sets. In order to apply this scheme to the problem of revealing granularity levels we compare distributions of term specificity of two data sets in a moving window. Points of maximum distance between these distributions are considered corresponding to the transition points between adjacent granularity levels.

1.3 Overview of the research report

The rest of this document is structured as follows. In Chapter 2, we overview some existing methods concerning ontology learning and evaluation. Chapter 3 is devoted to the problem of biomedical NER. In Chapter 4, we give some base knowledge about IMMSO and elaborate its stability for different characteristics of input data. Chapter 5 aims at presenting a method of revealing granularity levels of domain terminology that is accomplished using IMMSO as an optimization technique. Another useful application of IMMSO to the dialogue processing domain is described in Chapter 6. Finally, in Chapter 7, we draw our conclusions and future research work in ontology learning.

Chapter 2

Ontology learning and evaluation: state of the art

2.1 Ontology learning

There exist several paradigms of learning ontological relations. The most widespread of them are **linguistic patterns-based** and **clustering-based**.

The first group of approaches use lexico-syntactic patterns to capture a special type of relations. In general, the rule-based approaches show very high precision but very low recall, because they are learned to extract only explicit data from texts. Despite mentioned drawbacks, many researchers exploit this paradigm because it allows to extract more reliable types of relations. Hearst [20] who states at the beginning of this approach developed linguistic patterns to derive hypernym/hyponym relations. Berland et al. [9] worked on discovering meronyms. Poesio et al. [37] reported about possibilities to improve accuracy of extracted relations applying algorithms for anaphora resolution.

The second group of approaches are clustering approaches based on the distributional hypothesis (Harris [19]). The idea of this hypothesis consists in the assumption that meanings of words are defined by their contexts. Therefore, two words are similar if they share similar contexts. Within the limits of this hypothesis many methods were proposed. The majority of them contains the following elements:

1. Representation of a word through a vector of its attributes;
2. Definition of distance/similarity metrics;
3. Application of some clustering method.

Context of word can be captured using contextual window, but many successful works were accomplished exploiting syntactic dependencies between words. It is logic because knowledge of syntactic relations is an additional information, which allows to consider only the most informative words of a context (nouns and verbs) and compare them with the same syntactic elements of another context. The last point does the

comparison procedure more intelligible.

One of the first works dedicated to clustering of nouns is a work of Hindle [21]. According to Hindle, nouns are thought similar if they participate in similar verb frames. To construct the similarity measure he considered only subject and object syntactic relations and exploited mutual information of two words to calculate their similarity. Unfortunately, Hindle did not apply any technique to organize nouns into hierarchy.

Pereira et al. [36] proposed to measure distance between nouns through the relative entropy of their syntactic context distributions. Although the unique syntactic relation they exploited in the work was direct object relation.

Faure et al. [15] used conceptual clustering to construct concepts hierarchy. They extracted subcategorized verb frames whose arguments were nouns to be classified and in each phase of a clustering procedure merged the most similar verb frames. Their similarity measure was based on a number of intersections between frames arguments and on frequencies of these arguments. This method is semi-automatic, the user can influence on the process of frames generalization to avoid joining erroneous clusters.

The work of Caraballo [8] is interesting because she tried to join two paradigms of learning ontological relations. On the first step, she constructed unlabeled hierarchy of nouns using bottom-up clustering methods. Noun contexts were represented by their conjunctions and appositions and as similarity measure a cosine measure was used. On the second step, a corpus parsing using Hearst patterns was carried out in order to find hypernyms. For each cluster, a most frequent hypernym was chosen. Then, the obtained tree was compressed to delete useless concepts, which had not got any label. This approach although rather original suffer from serious drawbacks: the obtained taxonomy is binary and very redundant.

Bisson et al. [3] developed a workbench where they tried to unit and elaborate all existing experience on learning ontological relations based on distributional hypothesis. They implemented different distance metrics and gave a user to choose a large variety of syntactic dependencies for expressing a word context. Finally, they also provided their workbench with an estimation module, which allowed to evaluate recall and precision of the results using N extracted relations with the highest similarity score.

Recently important modification of clustering approaches which explore Formal Concept Analysis (FCA) emerged. FCA belongs to conceptual clustering algorithms where distances between objects are not represented through some quantitative value as in general clustering approaches but through comparing features of objects. More fully, the theory of FCA is based on the following notions: “formal context” and “formal concept”. The former denotes a set of features bijectively correlated with its set of objects. The latter identifies such a correlated pair - features-objects. It is important that a correspondence is bijective, which means that there is no object in a formal concept that does not share all the features of the formal context, and vice versa, there is no feature, which does not belong to all objects. Therefore, FCA

organizes a space of pairs (object, features) into a lattice of formal concepts from the most specific (i.e. those, which share maximum number of features) to the most general ones (i.e. those, which have only one feature in common).

FCA was first applied to the task of ontology learning by Cimiano et al. [11]. In order to construct pairs (object, features) they exploited syntactic dependencies in a sentence, namely, subject-verb, verb-object and verb-PP complement. Nouns were considered to be the objects and verbs played a role of their features. Cimiano et al. proposed a very interesting idea to regard verbs-like properties as generalizations of their corresponding set of nouns. This method sometimes gives more natural relations between concepts. In Cimiano et al. [10] a deep comparison of this approach with agglomerative and divisive clustering algorithms was realized. Authors evaluated constructed ontologies on a basis of a gold standard (see Section 2.2) exploiting criteria of effectiveness, efficiency and traceability. Although FCA in a worse case has an exponential time complexity, ontologies based on this approach provide a higher level of traceability.

2.2 Ontology evaluation

Due to great increase of available in Web ontologies, ontology evaluation task has become very important during the last decade. Although many researchers dedicated to resolving this task the problem of ontology evaluation still remains hardly realizable and vague.

The main approaches for evaluating ontologies that emerged during the last years can be classified into 5 following groups:

1. Approach based on a gold standard. Its main purpose is developing methods and criteria to compare 2 distinct conceptual graphs. Among the well-known works in this area we can mention a work of Maedche and Staab [28], which was the first attempt to formalize and automatize an evaluation procedure.

2. Data-driven approach, which deployed domain corpus in order to compare semantic content provided by ontology and raw texts. The most considerable works here are works of Brewster et al. [6] and Spyns [45].

3. Ontology validation through its metaproperties. The founders of this approach, Guarino and Welty [18], proposed to use philosophical notions of rigidity, identity and unity for ontological concepts in order to reveal the cases of ambiguity or inconsistency.

4. Application-based or task-based approach, which aims to assess ontologies through their application to some special task (Porzel et al. [38]).

5. Approach exploited a set of predefined quality criteria or metrics. This suite of metrics is manually elaborated by a specialist in order to estimate ontologies over these criteria with a final quantitative result for each ontology at the output. A well-known system that accomplished this assessment is OntoMetric worked up by Gomez-Perez

[27].

We would like to notice that first 3 groups of approaches mostly attempt to evaluate ontology accuracy, completeness, inconsistency, etc., while the fourth group is oriented to evaluate ontology goodness for a given application. Finally, the fifth group of methods aims at complete assessment of ontology.

At the same time, ontology evaluation could differ by the level of assessment. On the first level (data layer) only lexical entries are evaluated, on the second level (conceptual layer) also relations between concepts are taken into account, at last, on the context level the fitness of ontology to special task or application is assessed.

Therefore, we have 2 groups of evaluation principles, which, although dependent, assess ontologies from different aspects. In our review of existing evaluation methods, we will take into consideration both principles although, for regularity, we organize our survey by the first group of approaches.

Gold standard-based approach. The simplest method of ontology evaluation at a concept layer is one exploited such well-known Information Extraction measures as recall and precision. Precision measures a number of correctly identified items as a percentage of all identified items while recall measures a number of correctly identified items as a percentage of all correct terms. Correctly identified terms are also named as true positives, incorrectly identified terms as false positives and correct but not identified terms as false negatives. Using this denotements recall (R) and precision (P) can be defined as follows:

$$P = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$R = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (2.1)$$

The problem of correspondence between concepts of distinct ontologies is a difference in lexical entries that refer to the same meaning. Similar concepts could match partially or have absolutely different writing forms. If a partial similarity is taken into account in calculating precision and recall measures the formulas (2.1) can be modified in a following way [14]:

$$P = \frac{\text{true positives} + 1/2 \text{ partial}}{\text{true positives} + \text{false positives} + 1/2 \text{ partial}}$$

$$R = \frac{\text{true positives} + 1/2 \text{ partial}}{\text{true positives} + \text{false negatives} + 1/2 \text{ partial}} \quad (2.2)$$

For getting a tradeoff between both measures the F-measure is normally used, which is a harmonic mean of the two:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 R + P} \quad (2.3)$$

Recall and precision mostly measure a lexical similarity between two ontologies, they are not able to evaluate a semantic similarity. The work of Maedche and Staab [28] attempts to realize a comparison of two ontologies both at a data and conceptual level. In order to correlate lexical entries of two ontologies the authors propose to use Levenstein distance that it is an edit measure between two strings. It is equal to a minimum number of deletions, insertions and substitutions needed for converting one string to another one. For comparing ontologies as conceptual graphs Maedche and Staab introduce a notion of a semantic cotopy that is defined for each concept and represents a set of all its super- and subconcepts. Using this notion they introduce formulas for measuring similarity of hierarchical and no-hierarchical relations through taxonomic overlap and relational overlap. This work is one of the first works where ontology evaluation was accomplished in a formal way.

In practical tasks, where a user has to choose a most suitable ontology in order to incorporate it into his/her system a golden standard-based methods seem to be rather useless. Really, if he knew the golden standard he would choose it without any comparison. However, this approach might be very efficient for comparing different ontology learning methodologies.

Data-driven approach. This approach seems to us the most promising way to evaluate ontologies although previous corpus estimation has to be accomplished. At the moment, few works are dedicated to developing possible methodologies of this evaluation and the results still remain unsatisfactory or do not reported either. The first attempt was done by Brewster et al. [6] who applied a probabilistic method, namely, maximization of a conditional probability of an ontology O given a corpus C :

$$O^* = \operatorname{argmax}_O P(O|C) = \operatorname{argmax}_O \frac{P(C|O)P(O)}{P(C)} \quad (2.4)$$

A member $P(C|O)$ can be estimated by means of comparison between ontological concepts and relevant corpus terms but as the authors said the most common scenario in this case is one where there are items absent and unneeded. Brewster et al. proposed another methodology. First, they retrieve relevant terms from the corpus using Latent Semantic Analysis. Then, a clustering procedure deployed Expectation Maximization method is accomplished in order to find hidden topics with which extracted terms are related. At the same time, for each concepts of an ontology under evaluation a set of connected concepts are elicited using two levels of hypernyms of Wordnet. Finally, topic clusters and expanded ontology terms are used in (2.4) to measure similarity between the corpus and the ontology. The authors affirm that this method enables to measure both data level and conceptual level of similarity comparing not only term clusters and ontology concepts but also correspondence of ontological relations. An evident drawback of this method is that it is difficult to guess about a type of relations between obtained term clusters, and, therefore, ontology evaluating at an ontology level seems to be rather complicated and obscure.

Besides, the authors do not present any experiments carried out using their method. Therefore, it is impossible to speak about its effectiveness.

Spyns et al. [45] investigate a problem of whether it is feasible, at the moment, to accomplish correct ontology evaluation by means of raw data. In order to check it they analyze specificity, precision and recall of extracting from corpus triples. The gold standard here is determined by opinions of two independent experts. They apply z-stastic to select relevant words measuring a difference between a word frequency in a domain corpus and in a general one. Relevant triples are chosen due to a predefined threshold of lexical overlap between extracted triples and triples of the reference corpus. The experiments are realized with different confidence levels for terms extraction and with different thresholds for discovering relevant triples. The obtained results seem to be rather modest: when the specificity (which is a number of true negatives) and the recall (a number of true positives) exceed a level of 0.5 the precision only reaches 0.208. Therefore, the accuracy is not appropriate yet for ontology evaluation although this method can be used for ontology learning.

Application-based approach. This group of approaches associates ontology quality as its goodness for some application. As a result, this assessment procedure has no aim to discover the best ontology in general but one that suits best for some specific task. This evaluation seems to us rather logic because normally ontologies are used to be evaluated for their further reuse that implies their incorporation into another systems. However, in literature, this approach does not have much support. Among its drawbacks the following aspects are mentioned ([5]): (1) it is difficult to conclude about a quality of ontology in general; (2) ontology could be a small part of the application and, therefore, its effectiveness could be small and indirect; (3) comparison of ontologies must be accomplished under the same conditions, i.e. incorporating them into the same application, which is difficult to carry out due to different ontology formats.

Criteria-based approach. At the heart of the methods under these approach a suite of manually developed metrics associated with an ontology quality is lied. Burton-Jones et al. [7], considering an ontology quality from 4 aspects (syntactic, semantic, pragmatic and social) elaborated 10 metrics of ontology quality, among them consistency, clarity, accuracy, relevance, etc. Within their framework an automatic assessment of ontologies over the suite of metrics is carried out. A user can give weights to metrics augmenting important for his task and lowering insignificant. At the output of the evaluation a qualitative value is assigned to each ontology.

The OntoMetric system developed by Gomez-Perez [27] is based on the Analythic Hierarchy Process that is a multicriteria decision method. This system focuses mostly on evaluating ontologies with the purpose of their incorporation into another systems. The authors do not only point out quality aspects concerning their contents but also consider such important at implementation aspects as language, in which it was created, methodology that was exploited to construct it, software environments where it was builded and costs of using it inside of the system. Each of these quality aspects

called dimensions consists of a set of factors and their characteristics. OntoMetric contains 160 different characteristics represented as a multilevel tree. All these characteristics must be established by user. The evident drawback of this system is a need of user manual assessment of ontologies, which is complicated and time consuming task. The OntoMetric system attempts to accomplish complete ontology evaluating involving 3 mentioned above layers of assessment: data, ontology and application.

In this group of methods, a work by Orme et al. [35] can also be mentioned. Instead of establishing quality characteristics and methods of their calculation as in above methods, Orme et al. carried out a study of dependency of such ontology characteristics as complexity and cohesion on conceptual graph attributes. Among the attributes they used are average properties per class, average fanout per class, maximum depth tree, etc. To accomplish their study a set of ontologies was evaluated by experts in order to obtain objective values of complexity and cohesion and, then, a correlation between conceptual graph attributes and the manual estimations was realized using Pearson test. The authors obtained rather evident correlations, for example, between a number of leafs and complexity and cohesion of ontology. They also tried to evaluate an ontology in evolution that can be represented by characteristics of stability and completeness but no confident correlation was discovered.

Validation through metaproperties. This approach was developed by Guarino and Welty [18] and was implemented in the OntoClean system. They elaborated a formal theory of ontology evaluation through introducing philosophical notations. Guarino and Welty suggested to provide each concept and relation with such metaproperties as rigidity, identity and unity, which would help to discover inconsistency or ambiguity existing in ontology. The main disadvantage of this approach that a user need to add manually all the metaproperties for concepts of ontology and it is rather tedious and difficult.

Chapter 3

Biomedical Named Entity Recognition

3.1 Motivation

Recently the molecular biology domain has been getting a massive growth due to many discoveries that have been made during the last years and due to a great interest to know more about the origin, structure and functions of living systems. It causes to appear every year a great deal of articles where scientific groups describe their experiments and report about their achievements.

Nowadays the largest biomedical database resource is MEDLINE that contains more than 14 millions of articles of the world's biomedical journal literature and this amount is constantly increasing - about 1,500 new records per day [12]. To deal with such an enormous quantity of biomedical texts different biomedical resources as databases and ontologies have been created.

Actually NER is the first step to order and structure all the existing domain information. In molecular biology it is used to identify within the text which words or phrases refer to biomedical entities, and then to classify them into relevant biomedical concept classes.

Although NER in molecular biology domain has been receiving attention by many researchers for a decade, the task remains very challenging and the results achieved in this area are much poorer than in the newswire one.

The principal factors that have made the biomedical NER task difficult can be described as follows [49]:

(i) *Different spelling forms existing for one entity* (e.g. “N-acetylcysteine”, “N-acetyl-cysteine”, “NacetylCysteine”).

(ii) *Very long descriptive names*. For example, in the Genia corpus (which will be described in Section 3.3.1) the significant part of entities has length from 1 to 7.

(iii) *Term share*. Sometimes two entities share the same words that usually are headnouns (e.g. “T and B cell lines”).

(iv) *Cascaded entity problem*. There exist many cases when one entity appears inside another one (e.g. $\langle PROTEIN \rangle \langle DNA \rangle kappa3 \langle /DNA \rangle bindingfactor \langle /PROTEIN \rangle$) that lead to certain difficulties in a true entity identification.

(v) *Abbreviations*, that are widely used to shorten entity names, create problems of its correct classification because they carry less information and appear less times than the full forms.

This work aims to investigate and compare a performance of two popular Natural Language Processing (NLP) approaches: HMMs and CRFs in terms of their application to the biomedical NER task. All the experiments have been realized using a JNLPBA version of Genia corpus [24].

HMMs [39] are generative models that proved to be very successful in a variety of sequence labeling tasks as Speech recognition, POS tagging, chunking, NER, etc.[32, 50]. Its purpose is to maximize the joint probability of paired observation and label sequences. If, besides a word, its context or another features are taken into account the problem might become intractable. Therefore, traditional HMMs assume an independence of each word from its context that is, evidently, a rather strict supposition and it is contrary to the fact. In spite of these shortcomings the HMM approach offers a number of advantages such as a simplicity, a quick learning and also a global maximization of the joint probability over the whole observation and label sequences. The last statement means that the decision of the best sequence of labels is made after the complete analysis of an input sequence.

CRFs [26] is a rather modern approach that has already become very popular for a great amount of NLP tasks due to its remarkable characteristics [42, 30, 41]. CRFs are indirected graphical models which belong to the discriminative class of models. The principal difference of this approach with respect to the HMM one is that it maximizes a conditional probability of labels given an observation sequence. This conditional assumption makes it easy to represent any additional feature that a researcher could consider useful, but, at the same time, it automatically gets rid of the property of HMMs that any observation sequence may be generated.

This chapter is organized as follows. In Section 3.2 a brief review of the theory of HMMs and CRFs is introduced. In Section 3.3 different strategies of building our HMM-based and CRF-based models are presented. Since corpus characteristics have a great influence on the performance of any supervised machine-learning model the first part of Section 3.3 is dedicated to a description of the corpus used in our work. In Section 3.4 the performances of the constructed models are compared. Finally, in Section 3.5 we draw our conclusions and discuss the future work.

3.2 HMMs and CRFs in sequence labeling tasks

Let $\mathbf{x} = (x_1x_2\dots x_n)$ be an observation sequence of words of length n . Let \mathbf{S} be a set of states of a finite state machine each of which corresponds to a biomedical entity tag $t \in T$. We denote as $\mathbf{s} = (s_1s_2\dots s_n)$ a sequence of states that provides for our word sequence \mathbf{x} some biomedical entity annotation $\mathbf{t} = (t_1t_2\dots t_n)$.

HMM-based classifier belongs to naive Bayes classifiers which are founded on a joint probability maximization of observation and label sequences:

$$P(\mathbf{s}, \mathbf{x}) = P(\mathbf{x}|\mathbf{s})P(\mathbf{s})$$

In order to provide a tractability of the model traditional HMM makes two simplifications. First, it supposes that each state s_i only depends on a previous one s_{i-1} . This property of stochastic sequences is also called a Markov property. Second, it assumes that each observation word x_i only depends on the current state s_i . With these two assumptions the joint probability of a state sequence \mathbf{s} with observation sequence \mathbf{x} can be represented as follows:

$$P(\mathbf{s}, \mathbf{x}) = \prod_{i=1}^n P(x_i|s_i)P(s_i|s_{i-1}) \quad (3.1)$$

Therefore, the training procedure is quite simple for HMM approach, there must be evaluated three probability distributions:

- (1) initial probabilities $P_0(s_i) = P(s_i|s_0)$ to begin from a state i ;
- (2) transition probabilities $P(s_i|s_{i-1})$ to pass from a state s_{i-1} to a state s_i ;
- (3) observation probabilities $P(x_i|s_i)$ of an appearance of a word x_i in a position s_i .

All these probabilities may be easily calculated using a training corpus.

The equation (3.1) describes a traditional HMM classifier of the first order. If a dependence of each state on two preceding ones is assumed a HMM classifier of the second order will be obtained:

$$P(\mathbf{s}, \mathbf{x}) = \prod_{i=1}^n P(x_i|s_i)P(s_i|s_{i-1}, s_{i-2}) \quad (3.2)$$

CRFs are undirected graphical models. Although they are very similar to HMMs they have a different nature. The principal distinction consists in the fact that CRFs are discriminative models which are trained to maximize the conditional probability of observation and state sequences $P(\mathbf{s}|\mathbf{x})$. This leads to a great diminution of a number of possible combinations between observation word features and their labels and, therefore, it makes possible to represent much additional knowledge in the

model. In this approach the conditional probability distribution is represented as a multiplication of feature functions exponents:

$$P_{\theta}(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_0} \exp \left(\sum_{i=1}^n \sum_{k=1}^m \lambda_k f_k(s_{i-1}, s_i, \mathbf{x}) + \sum_{i=1}^n \sum_{k=1}^m \mu_k g_k(s_i, \mathbf{x}) \right) \quad (3.3)$$

where Z_0 is a normalization factor of all state sequences, $f_k(s_{i-1}, s_i, \mathbf{x})$, $g_k(s_i, \mathbf{x})$ are feature functions and λ_k, μ_k are learning weights of each feature function. Although, in general, feature functions can belong to any family of functions, we consider the simplest case of binary functions.

Comparing equations (3.1) and (3.3) there may be seen a strong relation between HMM and CRF approaches: feature functions f_k together with its weights λ_k are some analogs of transition probabilities in HMMs while functions $\mu_k f_k$ are observation probability analogs. But in contrast to the HMMs, the feature functions of CRFs may not only depend on the word itself but on any word feature, which is incorporated into the model. Moreover, transition feature functions may also take into account both a word and its features as, for instance, a word context.

A training procedure of the CRF approach consists in the weight evaluation in order to maximize a conditional log likelihood of annotated sequences for some training data set $D = (\mathbf{x}, \mathbf{t})^{(1)}, (\mathbf{x}, \mathbf{t})^{(2)}, \dots, (\mathbf{x}, \mathbf{t})^{(|D|)}$

$$L(\theta) = \sum_{j=1}^{|D|} \log P_{\theta}(\mathbf{t}^{(j)}|\mathbf{x}^{(j)})$$

We have used CRF++ open source ¹ which implemented a quasi-Newton algorithm called LBFGS for the training procedure.

3.3 Biomedical NE recognizers description

Biomedical NER task consists in detection of biomedical entities in a raw text and assigning them to one of the existing entity classes. In this section the two biomedical NE recognizers, we constructed, based on the HMM and CRF approaches will be described.

3.3.1 JNLPBA corpus

Any supervised machine-based model depends on a corpus that has been used to train it. The greater and the more complete the training corpus is, the more

¹<http://www.chasen.org/taku/software/CRF++/>

precise the model will be and, therefore, the better results can be achieved. At the moment the largest and, therefore, the most popular biomedical annotated corpus is Genia corpus v. 3.02 which contains 2,000 abstracts from the MEDLINE collection annotated with 36 biomedical entity classes. To construct our model we have used its JNLPBA version that was applied in the JNLPBA workshop in 2004 [24]. In Table 3.1 the main characteristics of the JNLPBA training and test corpora are illustrated.

Table 3.1: JNLPBA corpus characteristics

Characteristics	Training corpus	Test corpus
Number of abstracts	2,000	404
Number of sentences	18,546	3,856
Number of words	492,551	101,039
Number of biomed. tags	109,588	19,392
Size of vocabulary	22,054	9,623
Years of publication	1990-1999	1978-2001

The JNLPBA corpus is annotated with 5 classes of biomedical entities: protein, RNA, DNA, cell type and cell line. Biomedical entities are tagged using the IOB2 notation that consists of 2 parts: the first part indicates whether the corresponding word appears at the beginning of an entity (tag B) or in the middle of it (tag I); the second part refers to the biomedical entity class the word belongs to. If the word does not belong to any entity class it is annotated as “O”. In Fig. 3.1 an extract of the JNLPBA corpus is presented in order to illustrate the corpus annotation. In Table 3.2 a tag distribution within the corpus is shown. It can be seen that the majority of words (about 80%) does not belong to any biomedical category. Furthermore, the biomedical entities themselves also have an irregular distribution: the most frequent class (protein) contains more than 10% of words, whereas the most rare one (RNA) only 0.5% of words. The tag irregularity may cause a confusion among different types of entities with a tendency for any word to be referred to the most numerous class.

Table 3.2: Entity tag distribution in the training corpus

Tag name	Protein	DNA	RNA	cell type	cell line	no-entity
Tag distr.%	11.2	5.1	0.5	3.1	2.3	77.8

IL-2	B-DNA
gene	I-DNA
expression	O
and	O
NF-kappa	B-protein
B	I-protein
activation	O
through	O
CD28	B-protein
requires	O
reactive	O
oxygen	O
production	O
by	O
5-lipoxygenase	B-protein
.	O

Figure 3.1: Example of the JNLPBA corpus annotation

3.3.2 Feature set

As it is rather difficult to represent in HMMs a rich set of features and in order to be able to compare HMM and CRF models under the same conditions we do not apply such commonly used features as orthographic or morphological ones. The only additional information we exploit are parts-of-speech (POS) tags.

The set of POS tags was supplied by the Genia Tagger². It is significant that this tagger was trained on the Genia corpus in order to provide better results in the biomedical texts annotation. As it has been shown by [50], the use of the POS tagger adapted to the biomedical task may greatly improve the performance of the NER system than the use of the tagger trained on any general corpus as, for instance, Penn TreeBank.

3.3.3 Two strategies of HMM- and CRF-based models' building

As we have already mentioned, CRFs and HMMs have principal differences and, therefore, distinct methodologies should be employed in order to construct the biomedical NE recognizers based on these models.

Due to their structure, HMMs cause certain inconveniences for feature set representation. The simplest way to add a new knowledge into the HMM model is to specialize its states. This strategy was previously applied to other NLP tasks, such as POS tagging, chunking or clause detection and proved to be very effective [32].

Thus, we employ this methodology for the construction of our HMM-based biomedical NE recognizer. States specialization leads to the increasing of a number of states and to adjusting each of them to certain categories of observations. In other words, the idea of specialization may be formulated as a splitting of states by means of additional features which in our case are POS tags.

²<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

In our HMM-based system, the specialization strategy using POS information serves both to provide an additional knowledge about entity boundaries and to diminish an entity class irregularity. As we have seen in Section 3.3.1, the majority of words in the corpus does not belong to any entity class. Such data irregularity can provoke errors, which are known as false negatives, and, therefore, may diminish the recall of the model. It means that many biomedical entities will be classified as non-entity. Besides, there also exists a non-uniform distribution among biomedical entity classes: e.g. class “protein” is more than 100 times larger than class “RNA” (see Table 3.2).

We construct three following models based on HMMs of the second order (3.2):

1. only the non-entity class has been splitted;
2. the non-entity class and two most numerous entity categories (protein and DNA) have been splitted;
3. all the entity classes have been splitted.

It may be observed that each following model includes the set of entity tags of the previous one. Thus, the last model has the greatest number of states.

Besides, we carry out various experiments with a different number of boundary tags, and we conclude that only adding two tags (E - end of an entity and S - a single word entity) to a standard set of boundary tags, supplied by the JNLPBA corpus annotation, can notably improve the performance of the HMM-based model.

Consequently, each entity tag of our models contains the following components:

1. entity class (protein, DNA, RNA, etc.);
2. entity boundary (B - beginning of an entity, I - inside of an entity, E - end of an entity, S - a single word entity);
3. POS information.

With respect to the CRF approach, the specialization strategy seems to be rather absurd, because it was exactly developed to be able to represent a rich set of features. Therefore, instead of increasing of the states number the greater quantity of feature functions corresponding to each word should be used. Our CRF-based NE recognizer along with the POS tags information also employes context features in a window of 5 words.

The key point that should be drawn attention to is the POS set used in the splitting procedure. We think that the whole set of POS is rather redundant and contributes neither to the system accuracy, no to its stability. In order to split a non-entity class, the distribution of its in-class POS tags has been analyzed (Fig. 3.2). We realize several experiments to choose the best set of POS tags. As a result,

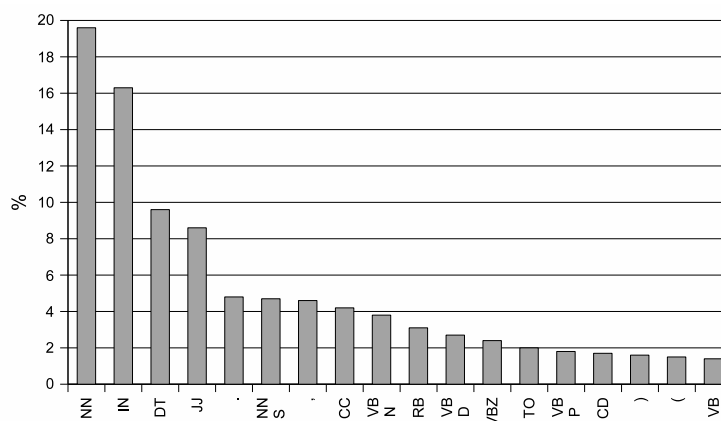


Figure 3.2: POS distribution inside of a no-entity category

the POS with a relative frequency of more than 1% is selected to participate in the entity tag balancing.

The classes of biomedical entities are divided according to the POS distribution within the class “Protein”. In order to participate in the splitting procedure, the most frequent POS tags are chosen (Table 3.3). As it may be noticed from Table 3.3, some parts-of-speech can appear only in certain parts of a biomedical entity (e.g. coma, brackets or conjunction never stay at the beginning of an entity).

Types of tags	POS
All	NN, JJ, NNS
I-tags	(, CC, “comma”
I- and E-tags	CD,)

Table 3.3: List of POS tags taken into account during the biomedical entity category splitting

3.4 Experiments and discussions

The standard evaluation metrics used for classification tasks are recall, precision and F-score introduced in (2.1),(2.3).

The first experiments we carry out are devoted to compare our three HMM-based models in order to analyze what entity class splitting provides the best performance. In Table 3.4, our baseline (i.e., the model without class balancing procedure) is compared with our three models. Although all our models improve the baseline, there is a significant difference between the first model and the other two models, which show rather similar results.

Table 3.4: Comparison of the influence of different sets of POS to the HMM-based system performance

Model	Tags number	Recall, %	Precision, %	F-score
Baseline	21	63.7	60.2	61.9
Model 1	40	68.4	61.4	64.7
Model 2	95	69.1	62.5	65.6
Model 3	135	69.4	62.4	65.7

In Table 3.5, the results we obtain with our CRF-based system are presented. Here, the baseline model takes into account only words and their context features. Model 1 is the final model, which uses also POS-tag information.

Table 3.5: The CRF-based system performance

Model	Recall, %	Precision, %	F-score
Baseline	61.9	72.2	66.7
Model 1	66.4	71.1	68.7

At first glance, if only the F-score values are compared, the CRF-based model outperforms the HMM-based one with a significant difference (3 points). However, when the recall and precision are compared their opposite behaviour may be noticed : for the HMM-based model the recall almost always is higher than the precision whereas for the CRF-based model the contrary is true.

In Tables 3.6, 3.7, recall and precision values of the detection of two biomedical entities “protein” and “cell type” for the HMM and the CRF approaches are presented. The analysis of these tables shows the higher effectiveness of HMMs in finding as many biomedical entities as possible and their failure in the correctness of this detection. CRFs are more foolproof models but, as a result, they commit a greater error of the second order: the omission of the correct entities.

Table 3.6: Recall values of a detection of “protein” and “cell type” for the HMM and the CRF models

Method	Protein	cell type
HMM	73.4	67.5
CRF	69.8	60.9

The certain advantage of the CRF model with respect to the HMM one could also be disputed by the fact that the best biomedical NER system [50] is principally based on the HMMs. Nevertheless, the comparison does not seem rather fair, because this system, besides exploiting a rich set of features, employs some deep knowledge resources and techniques such as biomedical databases (SwissProt and LocusLink)

Table 3.7: Precision values of a detection of “protein” and “cell type” for the HMM and the CRF models

Method	Protein	cell type
HMM	65.2	65.9
CRF	70.2	79.2

and a number of post-processing operations consisting of different heuristic rules in order to correct entity boundaries.

Summarizing the obtained results we can conclude that the possibility of an effective combination of CRFs and HMMs would be very beneficial. Since generative and discriminative models have different nature, it is intuitive, that their integration might allow to capture more information about the object under investigation. The example of a successful combination of these methods can be a Semi-Markov CRF approach which was developed by [40] and is a conditionally trained version of semi-Markov chains. This approach proved to obtain better results on some NER problems than CRFs.

3.5 Further work

In this chapter, we present two biomedical NE recognizers based on the HMM and CRF approaches. Both models are constructed with the use of the same additional information in order to compare fairly their performance under the same conditions. Since CRFs and HMMs belong to different families of classifiers two distinct strategies are applied to incorporate an additional knowledge into these models. For the former model a methodology of states specialization is used whereas for the latter one all additional information is presented in the feature functions of words.

The comparison of the results shows a better performance of the CRF approach if only F-scores of both models are compared. If also the recall and the precision are taken into account the advantage of one method with respect to another one does not seem so evident. In order to improve the results, a combination of both approaches could be very useful. As future work we plan to apply a Semi-Markov CRF approach for the biomedical NER model construction and also investigate another possibility of the CRF-based and the HMM-based models integration.

Chapter 4

IMMSO

In this chapter, we shortly describe principal characteristics of IMMSO and elaborate its stability with respect to different input parameters: a data volume, an inexactness of a model and a level of noise. We describe the experiments with polynomial models, and then compare the obtained results with those based on the traditional approach related to the approximation technique.

4.1 The method description

IMMSO belongs to the group of so-called evolutionary algorithms widely used in Data Mining. This method allows to determine the model of optimal complexity, which could describe/explain a given set of experimental data. Speaking 'model' we mean a formula, equation, algorithm, etc. At the moment, the IMMSO is not reflected well in the scientific literature in English, so we give its brief description and the conditions of its applications.

The IMMSO was suggested and developed by Ivakhnenko and his colleagues in 80s. This method does not require any a priori information concerning distribution of parameters of objects under consideration. Just for this reason the Ivakhnenko method proves to be very effective in the problems of Data Mining. Nevertheless it should be said that if such a priori information exists then other methods, for example, the methods of Pattern Recognition could provide essentially better results.

This method has one restriction: it cannot find the optimal model in any continuous class of models because its work is based on the competition of the models. That is why this method is titled as an inductive one. The main principle of model selection is the principle of stability: the models describing different subsets of a given data set must be similar. Here are the steps of the Ivakhnenko method:

- (1) An expert defines a sequence of models, from the simplest to more complex ones.

- (2) Experimental data are divided into two data sets: training data and control

data, either manually or using an automatic procedure.

(3) For a given kind of model, the best parameters are determined using, first, the training data and, then, the control one. For that any internal criteria of concordance between the model and the data may be used (e.g., the least squares criterion).

(4) Both models are compared on the basis of any external criteria, such as the criterion of regularity, or criterion of unbiasedness, etc. If this external criterion achieves a stable optimum, the process is finished; otherwise, more complex model is considered and the process is repeated from the step (3).

The method description contains the notions of 'internal criteria' and 'external criteria'. Let us mark the difference between them. The internal criteria use the same data both for determining model parameters and for evaluating model quality, whereas the external criteria deploy different data for these purposes. Usually the external criteria are constructed as non-negative functions with zero value in the best point.

IMMSO was successfully used in different natural-scientific and engineering applications [22], [23]. Recently it was used in natural language processing [1],[29]. In the following chapters we show its application to ontology learning and dialogue processing tasks.

4.2 Problem settings

IMMSO implies the fact the external criteria achieve a global extremum. Why do the authors of the method think that the external criteria reach an optimum? They suppose that experimental data contain: (a) a regular component defined by the chosen model structure and (b) a random component-noise. A simplified model does not react to the noise, but simultaneously it does not reflect the nature of objects. On the other hand, a sophisticated model can describe very well a real object behaviour but simultaneously such a model will also reflect a noise component. In both cases the differences between two models prove to be significant and the values of the penalty function (external criterion) are large. The principle of the model self-organization consists in the fact that an external criterion passes its minimum when the complexity of the model is gradually increased.

We should slightly correct this reasoning. Namely, if a chosen model does not reflect the behaviour of real objects or reflects it partially then the second component should not be named a noise. It is better to name it 'undefined component'. The contribution of this component to the experimental data can be comparative with the contribution of the model component. Such a circumstance distorts the results and minimum mentioned above may be located enough far from the real best point. Therefore, it is very important the user to choose correctly the model class. For this his/her choice should be based on the knowledge of object nature. Otherwise, the user should take enough generalized model class in order not to miss the real model

class.

Usually users mention the positive examples of the IMMSO applications. But the IMMSO can give essentially negative results if:

- (i) A volume of data is very limited;
- (ii) A chosen model class essentially differs from the real model;
- (iii) A level of external noise is high.

These circumstances can be commented by the following:

(i) Obviously, the limited data set does not permit to reveal models with high level of complexity. If the model complexity is associated with the number of its parameters then the number of data should be not less than the number of this model parameters. In reality, this condition is essentially more rigid (see, Section 4.3.1).

(ii) If the selected model class does not correspond the real object structure, then results prove to be unpredictable. As a rule simplified models are determined and this fact reflects the principle of stability (the consequence of the criterion of regularity, see Section 4.3.3).

(iii) If the level of external noise is high then the result proves to be dependent on its concrete realization. Usually the simplified models are created and this fact reflects the principle of independence from data (the consequence of the criterion of unbiasedness, see Section 4.3.3).

The purpose of the paper is to test the IMMSO sensibility with respect to a volume of data, an inexactness of the model class and a noise. In the paper, we consider polynomial functions. Obviously, the results will change if we deal with another classes of functions, for example, series of Fourier, etc. or with another type of models, such as differential models. Nevertheless, the qualitative conclusions will remain the same. The experiments were completed on the artificial data set as it was described in [22]. So, the results are easy interpreted.

In our work, we consider also the variant, when the model class and the level of noise are completely known. In this case, we use well-known Approximation Technique (AT). The results are compared with those obtained by the IMMSO.

4.3 Organization of experiments

4.3.1 Models under consideration

Firstly, we should fix the terminology. It concerns the notions of 'model', 'components of model', 'model complexity' and 'noise'. Speaking about a model in natural and technical sciences one implies the following:

- Mathematical model being presented by any equation or any system of equations;
- Algorithmic model being presented by a certain sequence of rules for data transformation;

In this paper, the mathematical models are considered in the following form:

$$\psi(x) = F(x) + \phi(x) + \omega(x), \quad (4.1)$$

where x is an independent variable defining the points of observations; $F(x)$ is a numerical function from a given class, reflecting essential (principal) components of a model; $\phi(x)$ is a numerical function, reflecting inessential (additional) components of a model; $\omega(x)$ is a random component (noise).

The following restrictions are accepted:

- Principal components are members of polynomial: $F(x) = a_0 + a_1x + a_2x^2 + \dots$
- Additional component is a periodical function $\phi(x) = b\cos(kx)$, where b is a parameter, defining its conclusion to the observation data; k defines a changeability of the additional component, which can be close or far from the changeability of the principal components.
- Noise $\omega(x)$ is a random function with zero mean and variance s^2 .

We want to note that the choice of polynomial and periodical functions as the model components is nonchance. Just these models are used in a time series analysis, related with season changes of temperature [22]. Coefficient $k \geq 5\pi$ for harmonic function could show itself on the interval of consideration [-1,1].

Elaborating the IMMSO we suppose that:

- additional component $\phi(x)$ is unknown; - standard deviation of noise s is unknown.

The IMMSO must determine the best polynomial in the form:

$$F(x) = a_0 + a_1x + a_2x^2 + \dots \quad (4.2)$$

It is well-known that the least square method, which is usually used in IMMSO, decreases the effect of noise in \sqrt{n} times if the number of data exceeds the number of model parameters in n times. So, in order to decrease the effect of noise at least in 2 or 3 times the number of observations should exceed the number of parameters in 5-10 times. Just these recommendations are usually given by the IMMSO developers [22], [23].

This circumstance defines the necessity to make several realization of noise component and then to generalize the results when we study the influence of noise to the performance of IMMSO and AT. It is important if we deal with small number of data and can not suppress the noise.

The term "Model complexity" usually denotes number and degree of relations of model components. In case of polynomial function the model complexity can be evaluated by highest polynomial degree.

If we deal with multivariable function then the volume of data very often proves to be insufficient to suppress data noise. For example, the complete polynomial of 3 variables of the 2-nd contains 10 members. If we have less than 100 observations then it happens to be impossible to reveal correctly such a simple model by means of

the typical combinatorial variant of the IMMSO (described above). In this case the other variant of IMMSO is used: it is so-called the Method of Grouped Arguments. It allows to make a model selection on each level of model complexity and to decrease the number of its parameters [22, 23]. In this paper we use only the mentioned combinatorial variant of the IMMSO.

4.3.2 Artificial data

Observation data are the values of function $G(x) = x^2 + b\cos(5\pi x)$ calculated in N points from the interval $[-1,1]$. Here: x^2 is considered as the principal component of a model, and $b\cos(5\pi x)$ as the additional one. In the experiments we consider the following values $b = \{0.1, 0.2, 0.5\}$. It is equal to 10%, 20% and 50% of the maximum value for x^2 on the interval $[-1,1]$.

The noise is added to these data, which is normal distributed random numbers with zero mean and root-mean-square deviation $s = \{0.1, 0.2, 0.5\}$. It also gives 10%, 20% and 50% of the maximum value for x^2 on the interval $[-1,1]$.

The number of points is equal to $N = 50, 1000$. Therefore, the training set and the control set contain 25 and 500 points respectively. Let us consider a complete polynomial 4.2 of the 2-nd order which evidently has 5 members. Then 5 and 100 points 'cover' each polynomial parameter respectively. It decreases the error of noise in $\sqrt{5} \sim 2$ and $\sqrt{100} = 10$ times. In case of small number of points ($N=50$) and high level of noise ($s=50\%$) we will perform 5 realization of a noise component and find the solution for each realization.

4.3.3 Result evaluation and energetic ratios

The correctness of solution is evaluated by the coincidence of polynomial order. That is if a polynomial of the 2-nd order is revealed (independently of the number of its members) then the method is accepted to work well. We mark once more that the concrete values of polynomial coefficients are not important for us: we reveal the optimal model structure but not the optimal values of its parameters under the revealed structure!

If it happens that the coefficient of the highest polynomial member is essentially less than the coefficient of the next member (approximately, 2 orders) than the model complexity decreases. For example, the polynomial of the 3 order $F(x) = 0.3 + 0.1x + 7x^2 + 0.05x^3$ on the interval $[-1,1]$ can be considered as a polynomial of the 2-nd order. But all such events will be marked. In all cases such a solution should be justified having in view, for example, a level of noise, etc.

Obviously, the polynomial function, periodical component and noise are the particular case of empirical models, which could be recovered with the IMMSO. In order to generalize the results it should describe the models under consideration with energetic ratios.

It is well-known that the power and the root-mean-square value of any function $f(x)$ defined on a given interval T can be calculated with the formulae:

$$W_f = \frac{\int_T f^2(x)dx}{T}, \quad U_f = \sqrt{W_f} \quad (4.3)$$

Thus, for the functions $F(x) = x^2$ and $\phi(x) = b\cos(kx)$ we have respectively: $U_F = 1/3, U_\phi = b/\sqrt{2}$.

The Table 4.3.3 contains the ratios $U_\phi/U_F, s/U_F$ and s/U_ϕ for the selected values of amplitude b (additional component) and selected values of root-mean-square deviation s (noise).

Table 4.1: Energetic ratios between model components

Ratio	b=0.1	b=0.1	b=0.1	b=0.2	b=0.2	b=0.2	b=0.5	b=0.5	b=0.5
	s=0.1	s=0.2	s=0.5	s=0.1	s=0.2	s=0.5	s=0.1	s=0.2	s=0.5
U_ϕ/U_F	0.2	0.2	0.2	0.4	0.4	0.4	1	1	1
s/U_F	0.3	0.6	1.5	0.3	0.6	1.5	0.3	0.6	1.5
s/U_ϕ	1.5	3	7.5	0.75	1.5	3.75	0.3	0.6	1.5

It is easy to see that we consider enough complex conditions of the experiments: the ratio of additional and principal components changes from 20% to 100%, and the ratio of noise and principal components changes from 30% to 150%.

4.3.4 Methods and criteria

1. Experiments with the IMMSO

The problem of IMMSO is to reveal polynomial function in a given form (2). Here the observation data (artificial data) are divided on training and control samples, the experiments are implemented, and the external criteria are calculated. The winner is defined on the basis of minimum of one or several external criteria described below.

The models, which are constructed on each step of IMMSO algorithm, use the Least Square Method (LSM). It should remind that we have two separate models: the first one based only on the training data set, and the second one based only on the control data set. Constructing these models LSM minimizes the variance of error between model data and data of observation:

$$\varepsilon^2 = \|F - D\|, \quad (4.4)$$

where $F = F_i$ is a vector of model function values in the points of observations x_i , $D = D_i$ is a vector of observation data.

There are many variants of the external criteria. Generally IMMSO uses the following two criteria:

- criterion of regularity K_r ;
- criterion of unbiasedness K_u .

Both criteria use the training data set and the test data set. The criterion of regularity reflects the difference between the model and the testing data, while the model is constructed on the training data set. Therefore, this criterion evaluates the stability of the model with respect to data variation. The criterion of unbiasedness reflects the difference between the two models-those constructed on the training and on the testing set, respectively. Thus, this criterion evaluates independence of the model from the data.

Different forms of these criteria can be proposed, a specific form depends on the problem. In our case we use these criteria in the following forms:

$$K_r = \frac{\sqrt{\sum_C (q_i(T) - q_i)^2}}{\sqrt{\sum_C (q_i)^2}}, \quad K_u = \frac{\sqrt{\sum_{T+C} (q_i(T) - q_i(C))^2}}{\sqrt{\sum_{T+C} (q_i)^2}}, \quad (4.5)$$

where T and C are the systems of equations 4.2 used for training and control, respectively; $q_i(T)$ and $q_i(C)$ are the "model" data that is the right part of equations with the parameters determined on the data of training and control, respectively; q_i are the experimental (artificial) data, i.e. the left part of the equations; i is the number of equation.

Sometimes a model can be better than another one according to the first criterion but worse according to the second one. Then a combined criterion is used:

$$K = \lambda K_r + (1 - \lambda) K_u, \quad (4.6)$$

where λ is a user-defined coefficient of preference. In our experiments, we use $\lambda = 2/3$, i.e. we consider the criterion of regularity as the main one.

2. Experiments with approximation technique

The problem of AT is also to approximate of the real model 4.1 by polynomial function 4.2. The model experiments are conducted on all data set without dividing on training and control sets. The model parameters are calculated with LSM using the criterion (4.4) as we did it above for the IMMSO.

The external criterion K_n is the difference between the normalized residual variance (4.4) and a given (known) variance of noise s^2 . Therefore, the external criterion can be formulated in the form:

$$K_n = |\varepsilon^2/N - s^2| \quad (4.7)$$

Such a criterion is really the external one because the value of s^2 does not related with data to be used for determining model parameters.

4.4 Experiments and results

For elaborating the IMMSO the following complete polynomial models are considered:

$$\begin{aligned} F_0(x) &= a_0 \\ F_1(x) &= a_0 + a_1x \\ F_2(x) &= a_0 + a_1x + a_2x^2 \\ F_3(x) &= a_0 + a_1x + a_2x^2 + a_3x^3 \end{aligned}$$

4.4.1 Stability with respect to a data volume

First, the dependency of results of a data volume was explored. In these experiments, we took the following parameters:

- the amplitude of the undefined component $b = 0.1$;
- the level of noise $s = 0.5$.

In Fig. 4.1, diagrams of combined external criteria for two different volumes of data $N=50$ and $N=1000$ are presented. It can be seen that little amount of data leads to the simplest model selection as an optimal one. However, for less data set we have more strongly pronounced minimum when the model complexity is equal to 2 than for the larger one. This fact demonstrates a greater possibility of the model complication if a large amount of data is used.

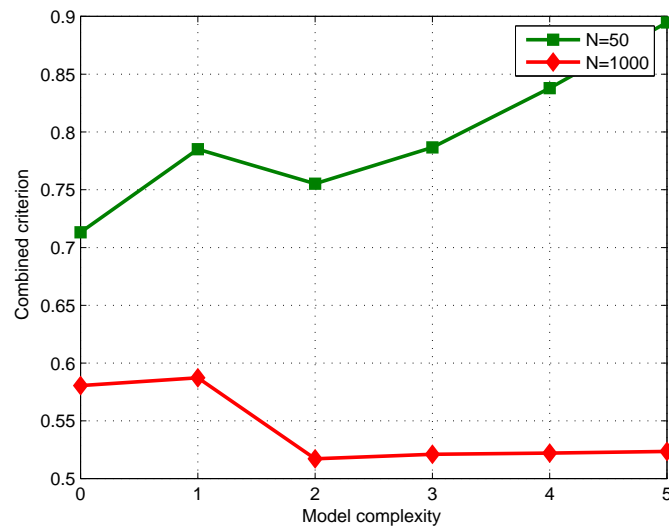


Figure 4.1: Behaviour of the external criterion for different volumes of data

4.4.2 Stability with respect to the unexactness of model

The aim of these experiments was to investigate an influence of the undefined component on the IMMSO stability. Another model parameters were fixed as:

- the level of noise $s = 0.1$;
- volume of data $N = 50$.

Fig. 4.2 shows diagrams of combined criteria for 3 parameters of the undefined component: $b=0.1$, $b=0.2$, $b=0.5$. It might be noticed that the stability of the IMMSO falls as a level of the additional component grows. Thus, the method tends to displace the solution towards more complex models.

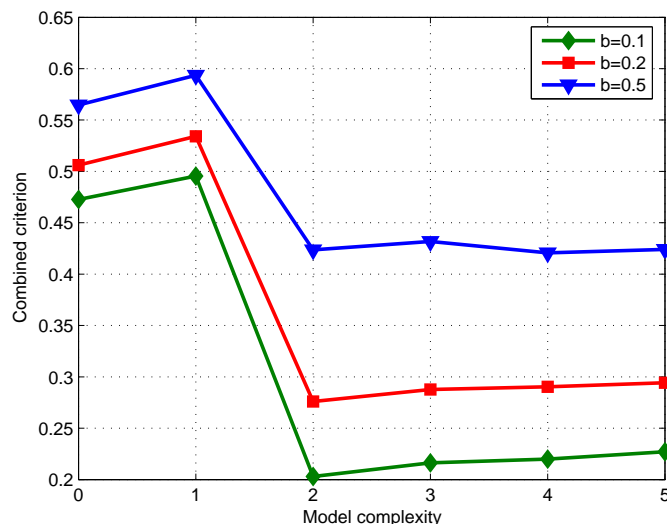


Figure 4.2: Behaviour of the external criterion with respect to the different amplitudes of the undefined component

4.4.3 Stability with respect to the noise

In order to evaluate an influence of the noise on the method solution we accomplished several experiments with a various level of noise: $s=0.1$, $s=0.2$, $s=0.5$ with another parameters equal to:

- the amplitude of the undefined component $b = 0.1$;
- volume of data $N = 50$.

Fig. 4.3 where diagrams of combined criteria for different levels of noise are represented shows a rather interesting results opposite to those, obtained for different amplitudes of the undefined component: stability falls with decreasing of the noise. This phenomena might be explained by the fact that the noise does not have such a regular high frequency character as the additional component and, therefore, for some

realizations of noise the model is capable to adjust its fluctuations. It must noticed that, nevertheless, the IMMSO method correctly finds a real model for all levels of noise, used in the experiments.

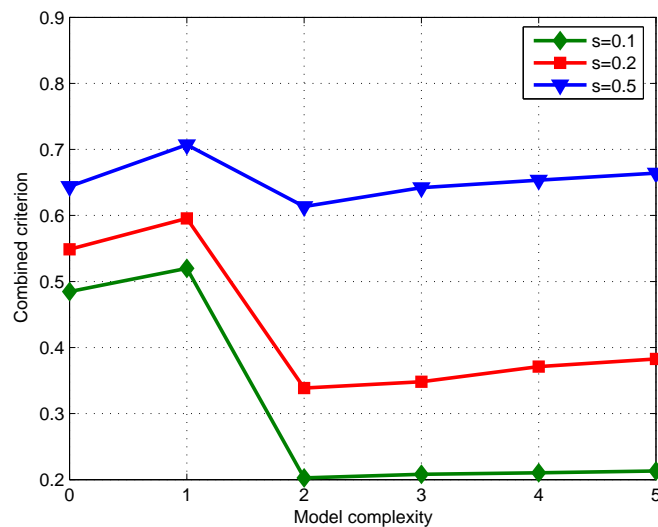


Figure 4.3: Behaviour of the external criterion with respect to the different levels of external noise

4.4.4 Model self-organization for different types of external criteria

It is interesting to analyse a contribution and tendency of each type of criteria. In order to estimate it we realized our experiments for low and high level of noise: $s=0.1$, $s=0.5$. All the rest parameters were as in above experiment. Analysing the obtained results (Fig. 4.4, 4.5) we can conclude that the criterion of regularity tends to approximate the real model by more complex models, while the criterion of unbiasedness always consider a trivial model as an optimal one. And, as we have already mentioned in the previous section, the combined criterion shows more stable results for a greater level of noise. This stability provides by the unbiasedness criterion behaviour.

4.4.5 Results with the Approximation Technique

The results obtained with the AT are shown in Fig. 4.6, 4.7. In order to compare its performance with the IMMSO we studied its for different data volumes $N=50$, $N=1000$ ($b=0.1$, $s=0.5$) (4.6) and with different levels of noise: $s=0.1$, $s=0.2$, $s=0.5$ ($N=50$, $b=0.1$) (4.7). The obtained results demonstrate an expressed tendency of this

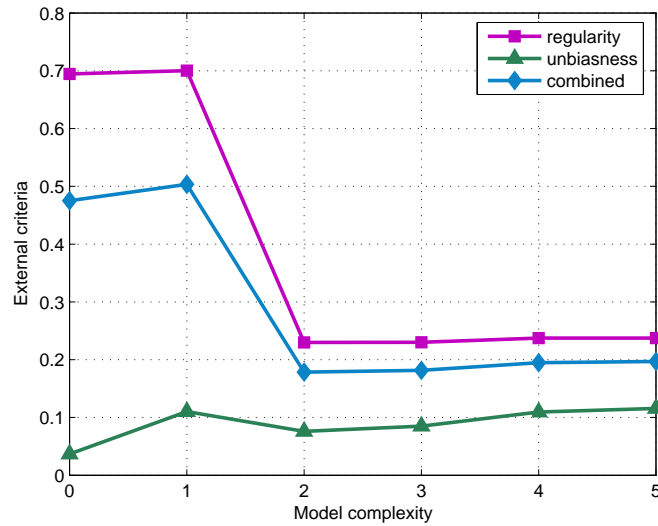


Figure 4.4: Behaviour of different types of external criteria for low levels of noise

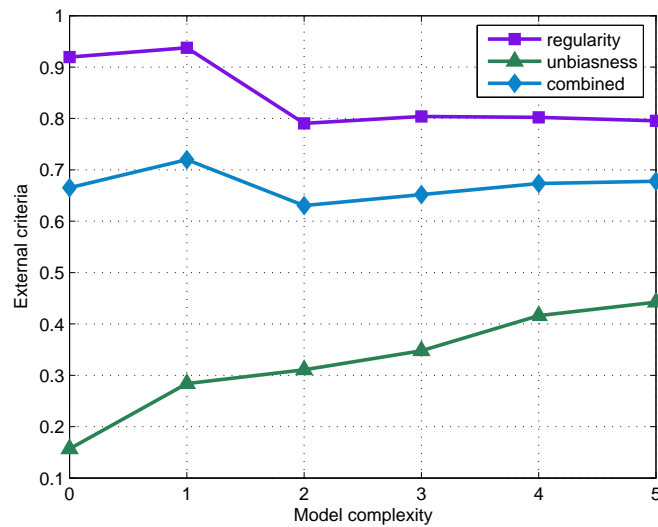


Figure 4.5: Behaviour of different types of external criteria for high levels of noise

technique to sophisticate a solution. Therefore, with the limits of this approach if even a little noise is present it is impossible to find the real model.

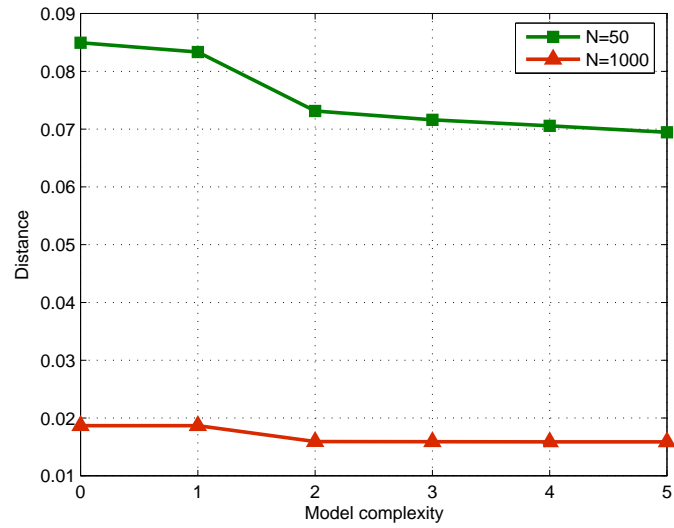


Figure 4.6: Results of approximation technique for different data volumes

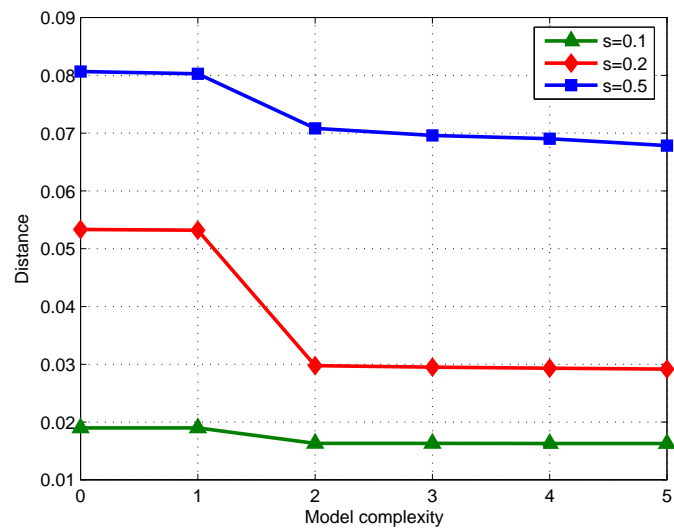


Figure 4.7: Results of approximation technique for different levels of external noise

Chapter 5

Revealing granularity of domain terminology

The aim of this work is to suggest a method of domain terms selection for different granularity levels. First, we give a definition of corpus-based term granularity and propose entropy-based and standard deviation-based weighting schemes for its evaluation. A chosen term weighting scheme is a decisive factor of granularity approximation quality. We declare a hypothesis of how to reveal boundaries between different granularity levels using our modified version of IMMSO. Although the suggested method demonstrates stability in the framework of our hypothesis some additional study of its reliability must be accomplished and other more precise weighting schemes should be applied.

5.1 Introduction

Evaluation of ontology granularity level keeps to be a difficult and weak-lighted in literature problem, although the importance of its solution is indisputable. A notion of granularity is used in a very intuitive way, neither its formal definition nor a mode of its measuring has been proposed up to now. Ontology granularity can be considered from different aspects: either in a lexical level, which refer to a granularity of ontology concepts or in a conceptual level where expressiveness of ontology properties and relations is in a center of investigation.

In this work, we consider a problem of ontology granularity in a lexical level (the lowest level of expressiveness according to *Ontology Summit 2007*¹). Therefore, we aim at evaluating only granularity of ontology concepts without taking into account other ontology components. In other words, approximation of ontology by a list of its concepts is realized. Although it is a very rude approximation we argue that it is a first step of ontology granularity evaluation.

¹<http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2007>

It is easy to observe that term granularity is related with a notion of term specificity/generality. Really, general terms can be considered as coarse-grained ones while very specific terms as fine-grained ones. It is widely-known that in Text Mining different term weighting schemes are used to measure term specificity, one of the most famous of them is inverse document frequency (IDF). Therefore, it seems logic to express term granularity through term specificity in order to convert a problem of revealing granularity levels into the problem of splitting the specificity scale into the segments with equal granularity.

What is meant by a notion of equal granularity? Mathematically it can be expressed that their values of specificity are very close. Immediately another question emerges: how close the term weights must be in order to be considered belonging to the same granularity level?

Attempting to answer the last question we apply to IMMSO. As we have no information about which granularity level is an optimal one because it completely depends on a task, moreover, as our aim is revealing various granularity levels and not only one of them, we develop a modified version of the IMMSO. With more detail the optimization scheme will be described in Section 5.4.

The the chapter is organized as follows. In Section 5.2 we define a problem of corpus-based term granularity and in Section 5.3 offer two different schemes of its approximation. In Section 5.4 we explain the application of IMMSO to the problem of revealing levels of granularity. The results of our experiments are described in Section 5.5. Finally, in Section 5.6 we summarize the obtained results and draw plans for future work.

5.2 Corpus-based term granularity

In this section, we will speak about a notion of corpus-based granularity of domain terms, i.e., about a granularity that can be calculated on a basis of a domain corpus. We wish to find a way of expressing granularity through some real scale, which would reflect characteristics of domain terms. Evidently, granularity can be considered in a scale of term specificity in view of the fact that more specific terms have finer-grained level and vice versa. The unique difference between term granularity and specificity, in our opinion, is that term granularity must be expressed through a discrete scale whereas this is not so necessary for specificity. Really, it is sufficient to know to which granularity level a given word belongs other addition information is redundant. In the light of the aforesaid, we propose a following definition of granularity levels:

Definition 1: Granularity levels are classes with close term specificity.

This definition means that granularity levels create a partition of a specificity range into the segments with close specificity and, accordingly, words with close specificity belong to the same granularity level.

Definition 2: Specificity range of domain corpus is a segment on a specificity

scale that cover all specificity weights of domain terms.

Evidently, specificity weights grow from general to more specific (rare) terms and normally a part of low frequent words in a corpus is rather great [13]. Some of them do not belong to the corpus domain and only introduce noise. Therefore, it is more convenient to work with an inverse specificity scale that gives lower weights to more specific terms and higher weights to more general ones. Henceforth, in our further reasoning we will imply this type of scale.

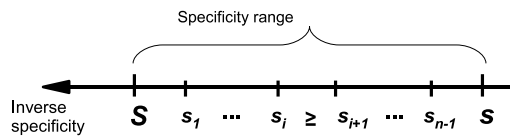


Figure 5.1: Inverse specificity scale

In Fig. 5.1 the inverse specificity scale is presented. It ranges from a value S corresponding to the lowest specificity and, therefore, the coarsest-grained granularity boundary to a value s , that refers to the highest specificity and the finest-grained granularity boundary. We define operations \geq and \leq in order to compare two specificity values. Let s_0, s_2, \dots, s_n be a partition of the inverse specificity scale into n segments, such that $s_i \geq s_{i+1}$ and $s_0 = S$ and $s_n = s$. We call the points s_i as transition points because they stay in boundaries between two different granularity levels. Using these denotations the problem of revealing granularity levels can be formulated in the following way:

Problem definition: Revealing granularity levels is equivalent to the problem of transition points collocation on the specificity range.

Thus, granularity problem can be splitted into two subproblems:

1. Approximation of specificity of domain terms by means of some term weighting scheme. Obviously, the results of this approximation are not only determined by a quality of applied scheme but also by a size of a document collection. The specificity weighting will be more precise if a corpus is rather large.

2. Collocation of transition points on a specificity range. It is carried out by an optimization method, a version of the IMMSO, that we elaborated specially for resolving this problem (See Section 5.4).

5.3 Specificity approximation

In the previous section, we expressed granularity through term specificity. In this section, we attempt to find appropriate ways of measuring term specificity. One of the most popular scheme is one based on IDF [44]. It relies on the fact that more common terms appear in a greater amount of documents whereas more specific ones normally

occur only in few documents. Let $D = \{d_1, \dots, d_i, \dots, d_N\}$ be a document collection under consideration. Inverse document frequency of a term w is represented as follows:

$$idf_w = \log \frac{N}{|\{d_i : w \in d_i\}|}. \quad (5.1)$$

The main disadvantage of such a measure is that it ignores the information about relative frequencies of a term in each document and, therefore, terms that appear in the same number of documents are given the same weight. For that reason we use alternative more precise measures of term specificity. One of them, named information-theoretic (or entropy-based) measure was firstly introduced by [47] and proved to be a generalization of the IDF measure. The other term weighting measure (standard deviation-based) exploits the standard deviation of term occurrences in the documents as a measure of its specificity.

5.3.1 Entropy-based specificity

Let us consider a number of occurrences of a term w in documents as a random variable. Given a document collection we obtain an evidence $X = (x_1, \dots, x_N)$ where x_i is a frequency of the word w in the document d_i . A probability distribution $P = (p_1, \dots, p_N)$ corresponding to the random variable X is defined as follows:

$$p_i = Pr(d_i|X) = \frac{x_i}{\sum_{i=1}^N x_i}, \quad i = 1, \dots, N \quad (5.2)$$

The entropy of a variable X is represented in the following way [43]:

$$H(X) = - \sum_{i=1}^N p_i \log(p_i) \quad (5.3)$$

According to the notion of Shannon entropy it is a measure of the uncertainty associated with a random variable. It is easy to see that it reaches a maximum value in a case of uniform distribution of a random variable X : $p_i = \frac{1}{N}$ for $i = 1, \dots, N$. Moreover, the entropy value decreases when the irregularity of probability distribution rises and it gets a value 0 when the “event” associated with this variable occurs only once. In our case, “event” refers to an appearance of a given term in a document. Therefore, the higher values of entropy correspond to more general terms and its lower values to more specific ones. This fact agrees with the inversion of the specificity scale. Summarizing all aforesaid, our first method of specificity approximation consists in:

Method 1: Term specificity is approximated by term entropy using the formula (5.3).

5.3.2 Standard deviation-based specificity

This weighting scheme also assumes that specificity can be determined by irregularity of term frequency distribution over a document collection. From statistics it is known that dispersion of a random variable can be expressed by its variance or standard deviation. Using the denotations introduced in the previous subsection a mean and a standard deviation of a random variable X can be presented as follows:

$$m = E(X) = \frac{1}{N} \sum_{i=1}^N x_i, \quad \sigma = \sqrt{E((X - m)^2)} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - m)^2} \quad (5.4)$$

The standard deviation must be normalized by the mean in order to bring distributions of all terms to the same mean for being able to compare dispersions of different terms. Obviously, normalized standard deviation and specificity are directly proportional and, therefore, according to the inverse specificity scale requirement, the weight of a term based on the standard deviation is represented as follows:

$$S(X) = \frac{m}{\sigma} \quad (5.5)$$

Thus, we can formulate another method of term specificity evaluation as:

Method 2: Term specificity is approximated by normalized standard deviation of term frequency over the whole document collection using the formula (5.5).

5.4 A method of detecting granularity levels

In Section 5.2, we established a relation between granularity and specificity of domain terms and formulated the problem of revealing granularity levels through the problem of partitioning the specificity range on the classes with close values of specificity. We also mentioned that this problem was reduced to the problem of transition points positioning. In this section, we suggest a way of detecting transition points with the use of an optimization method that can be considered as a version of IMMSO [23].

First of all, we want to make clear what is meant by a notion of 'model' and 'model complexity' in our case and to mark the main differences between our method and the IMMSO optimization scheme. A notion of 'model' implies a group of domain terms whose specificity is defined by some weighting scheme. Let us associate each domain term w with its inverse specificity value s_w and let us order all the words by decreasing of their inverse specificity. Evidently, all the terms will be collocated between maximum and minimum values of the specificity range. If we fix some threshold s^* of specificity in the specificity range a group of terms whose inverse specificity is greater than this threshold will be obtained (Fig. 5.2). Moving the

threshold to smaller values of inverse specificity, the group of extracted terms will grow because of adding finer-grained terms. As increasing a number of terms of the model implies a growth of the model complexity, by the model complexity we mean a specificity threshold.

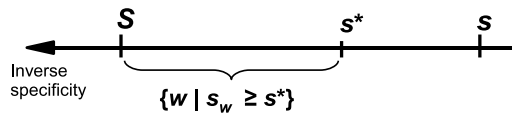


Figure 5.2: Term selection by specificity

According to its definition, IMSSO determines a model of optimal complexity or, in our case, a model of optimal granularity. As for each task there exists its own optimal granularity level, this problem cannot be resolved without some additional information about the task. However, revealing granularity levels can be thought as finding a model of optimal complexity in a local specificity window. Let us explain this strategy with more details.

IMMSO requires dividing a corpus into two parts: a training data set and a control data set. Obviously, specificity ranges of both data sets must be very close although they are not necessary the same. Let us fix the same window $\Delta s = [s_1, s_2]$ within the specificity ranges of training and control data sets and let us denote as $p_t = p_t(\Delta s)$ and $p_c = p_c(\Delta s)$ specificity distributions of terms covered by this window for both data sets respectively. Our hypothesis of transition points detecting consists in:

Hypothesis: If specificity distributions of training and control data sets over some specificity window Δs are close (have minimum distanced(p_t, p_c)) the terms contained in this window belong to the same granularity level. The windows where two distribution reach maximum distance $d(p_t, p_c)$ contain points of instability, i.e. the transition points from one granularity level to another one.

Ex hypothesi, we have to reveal points of maximum distance between specificity distributions of the training and the control data sets. It can be easily observed that the distance between two data set distributions corresponds to a notion of the external criterion in the terminology of IMMSO. As our two methods for weighting term specificity have different nature (Section 5.3) we use distinct external criteria for measuring distance between specificity distributions. We believe that term weighting method and external criterion should be consistent. For example, if for term weighting an information measure is used the same type of measure must be chosen for external criterion. This supposition makes us use a symmetrical variant of Kullback-Leibler distance [25] for entropy-based and a normalized version of Euclidean distance for the standard deviation-based weighting scheme. The definitions of these distances are presented below.

1. Relative entropy (or Kullback-Leibler distance) using above denotations is

formulated as follow:

$$K_1(\Delta s) = d(p_t, p_c) = \sum_{W_c(\Delta s)} p_t \log \frac{p_t}{p_c} + \sum_{W_t(\Delta s)} p_c \log \frac{p_c}{p_t}, \quad (5.6)$$

where $W_t(\Delta s)$ and $W_c(\Delta s)$ are groups of selecting terms in the training and control data sets respectively corresponding to the specificity window Δs .

2. Normalized version of Euclidean distance is defined as:

$$K_2(\Delta s) = d(p_t, p_c) = \sum_{W_c(\Delta s)} \frac{\sqrt{(p_t - p_c)^2}}{p_c} + \sum_{W_t(\Delta s)} \frac{\sqrt{(p_t - p_c)^2}}{p_t} \quad (5.7)$$

5.5 Experiments and results

5.5.1 Corpus characteristics

In our experiments, we use a corpus named hep-ex [33], originally stored in CERN. It consists of abstracts from particle physics domain. We carry out some preprocessing techniques before term weighting, namely, eliminating stop words and stemming with the Porter stemmer. We exploit neither syntactical parsing to select only principal parts of speech nor procedures for detecting compound words. Although all mentioned techniques can improve the quality of selected terms, the aim of our work was mostly not to obtain high-quality results but to analyze the possibility of granularity levels detection. The main characteristics of the hep-ex corpus after the preprocessing are done in Table 5.1.

Table 5.1: Hep-ex corpus characteristics

Size of the corpus(byte)	962,802
Number of abstracts	2,922
Total number of terms	135,969
Vocabulary size	6,150
Term average per abstract	46.53

5.5.2 Detecting levels of granularity

According to our modified version of IMMSO, granularity levels boundaries coincide with the maximum distance between term specificity distributions of two data sets. To reveal these points we calculate the external criterion function in a moving window. The shift of the moving window is equal to $0.01 * (S_t - s_t)$ and $0.01 * (S_c - s_c)$ for training and control data sets respectively, which corresponds to 1% from a specificity range. The length of the window is chosen experimentally. Evidently, very

short windows will give rather noisy and instable results, whereas long windows will smooth all distance peculiarities. In Fig.5.3-5.4 the experimental results with different window lengths (5%,7% and 9% of the specificity range) are shown. A window of length 9% is rather rude and smooths many extremums, which can be easily identified with other windows. Windows of 5% and 7% demonstrate similar results although in the case of entropy-based specificity some instability for high values of specificity appears. Therefore, we carry out all our experiments with the window length equal to 7%.

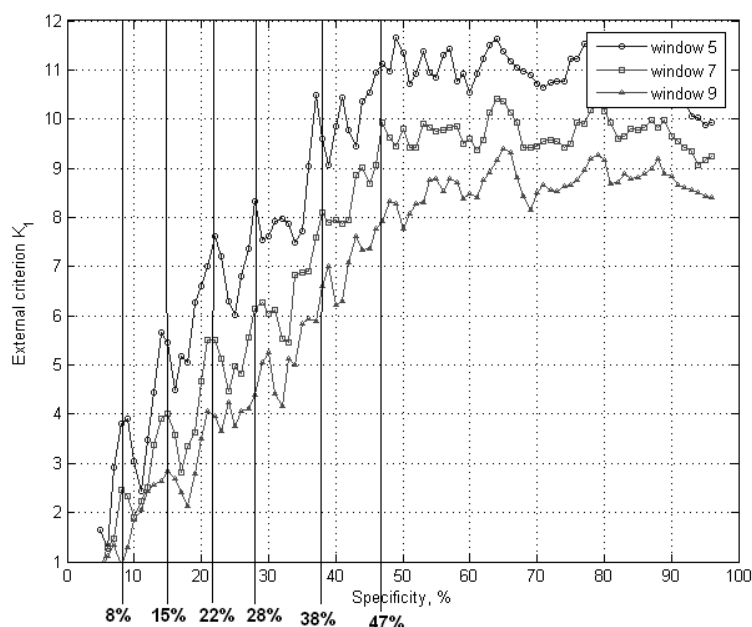
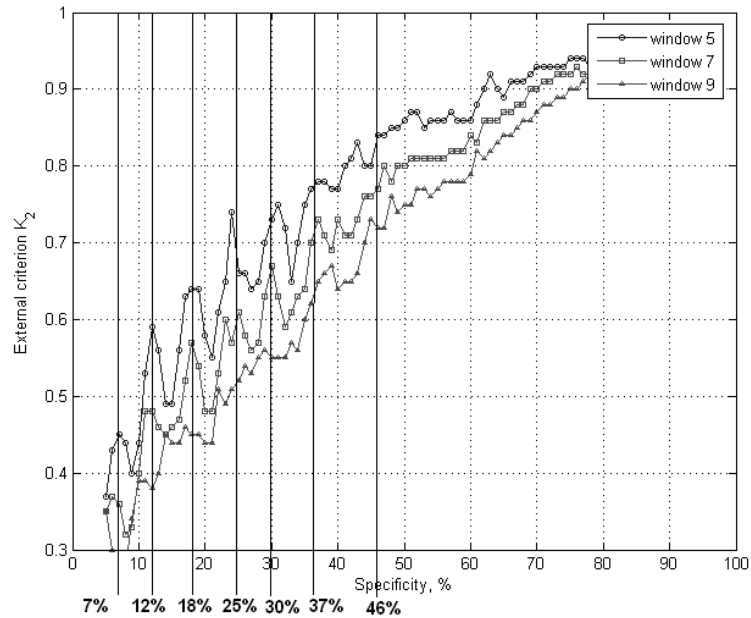
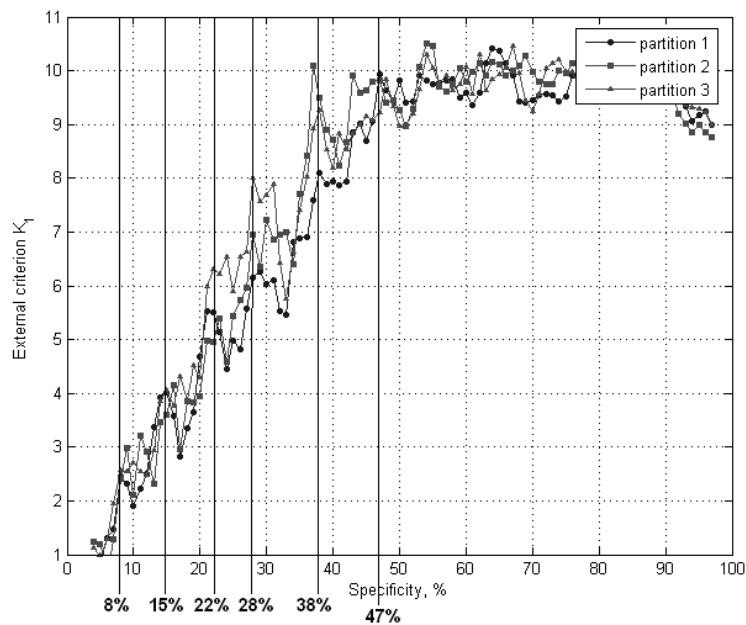


Figure 5.3: Behaviour of the K_1 criterion. Case of different window lengths

In order to demonstrate the stability of the revealed extremums, further experiments for different partitions of the document collection are realized (Fig.5.5-5.6). These figures show a good concordance and the obtained results can be considered as rather stable and reliable.

Analyzing Fig. 5.3-5.6 there can be seen several maximums of the external criteria: 6 for the criterion K_1 and 7 for the criterion K_2 . Some of the existing maximums, for example, those that are clearly seen on the entropy-based criterion K_1 for high values of specificity (Fig. 5.3), are supposed to be of poor reliability because they are not so considerable for other partitions (Fig. 5.5).

The unequal number of transition points obtained as the result of applying different specificity measures seems to us rather explainable because different ways of specificity approximations yield distinct specificity distributions of domain terms. Therefore, the method of specificity approximation is a decisive factor of the quality of the obtained results. In Table 5.2 a selected list of words ordered by the increase

Figure 5.4: Behaviour of the K_2 criterion. Case of different window lengthsFigure 5.5: Behaviour of the K_1 criterion. Case of different partitions of domain corpus

of term specificity for both weighting schemes is presented. As it can be seen there exists a slight difference in a term order although it is not very considerable. Table 5.3 reports the number of terms contained on the 3 first granularity levels. For the

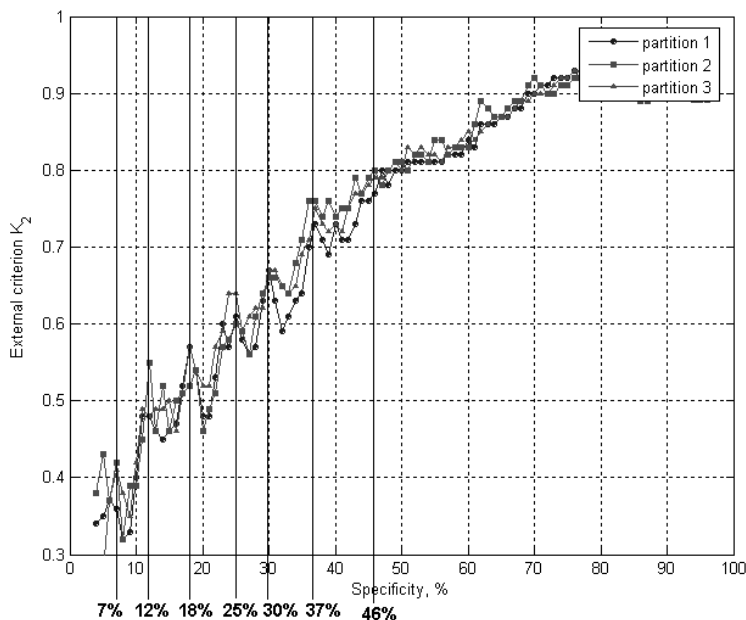


Figure 5.6: Behaviour of the K_2 criterion. Case of different partitions of domain corpus

entropy-based term weighting we have twice as much terms on the first granularity level. This can be explained by the fact that we miss the first extremum of the K_1 criterion due to relatively large window length for detecting such a small specificity value. If this assumption is true we must correlate the second level of the entropy-based granularity with the third level of the standard deviation-based one.

Table 5.2: Term list ordered by the increase of specificity

Entropy-based	Standard deviation-based
1. mesur	1. us
2. us	2. mesur
3. result	3. result
4. detector	4. present
5. data	5. detector
6. present	6. data
...	...
21. hadron	24. hadron
33. partiel	43. partiel
49. physic	47. physic
63. quark	71. electron
66. photon	78. photon
67. electron	81. quark
...	...

Table 5.3: Number of terms in each granularity level

Granularity level	Entropy-based	Standard deviation-based
1	94	40
2	96	44
3	107	67

Analyzing the obtained results we see two main drawbacks of our method:

1. **Weighting schemes for approximation specificity are not very precise** (Table 5.2). They tend to underestimate topic words that are in the center of consideration of a document collection. As a result, such a specific word as “hadron” has got lower value of specificity for both weighting schemes than words “particle” and “electron”. Therefore, in such a method realization (using the proposed specificity approximation) it is impossible to exploit obtained results for further ontology construction.

2. **The hypothesis introduced in Section 5.4 is needed to be verified.** Although for us this hypothesis seems rather logic, some verification procedures must be accomplished. We think that clustering of a document collection realized for each defined granularity level would help to verify our hypothesis, although gold standards of classification for different granularity levels are needful. It is expected that clustering results will better coincide with the gold standard ones if all the terms from the corresponding granularity level participate in clustering.

5.6 Summary and future work

Our work aims to formalize a notion of granularity and to suggest possible ways of its evaluation. We introduce a definition of the corpus-based term granularity through a notion of term specificity. Two different weighting schemes for the specificity approximation are suggested, one of them based on the term entropy and another one - on the variance of term frequency over a document collection. For revealing granularity levels a modified version of IMMSO is applied. The obtained results are evaluated on stability in the framework of our hypothesis asserted that the maximums of the criterion functions point to the granularity levels boundaries.

In future, we plan to apply and test other weighting schemes for specificity approximation, e.g. those used in text categorization [31],[48]. Also we are going to verify the reliability of our method by means of clustering.

Chapter 6

Constructing empirical models for automatic dialogue processing

Automatic classification of dialogues between clients and a service center needs a preliminary dialogue parameterization. Such a parameterization is usually faced with essential difficulties when we deal with politeness, competence, satisfaction, and other similar characteristics of clients. In the paper, we show how to avoid these difficulties using empirical formulae based on lexical-grammatical properties of a text. Such formulae are trained on given set of examples, which are evaluated manually by an expert(s) and the best formula is selected by the Ivakhnenko method of model self-organization. We test the suggested methodology on the real set of dialogues from Barcelona railway directory inquiries for estimation of passenger's politeness.

6.1 Problem setting

Nowadays, dialogue processing is widely used for constructing automatic dialogue systems and for improving service quality. By "dialogue" we mean a conversation between a client and a service center, and by "processing" we mean a classification of clients. Politeness, competence, satisfaction, etc. are very important characteristics for client classification but their formal estimation is quite difficult due to the high level of subjectivity. Thus, these characteristics usually are not taken into account or they are estimated manually [2].

In this work, we aim to construct an empirical formula to evaluate the mentioned characteristics, which is based on:

- (i) objective lexical-grammatical indicators related to a given characteristic;
- (ii) subjective expert opinion about dialogues.

The selection of lexical-grammatical indicators depends on expert experience. However, some simple indicators are often obvious, e.g. polite words for estimation of politeness, "if-then" expressions for the estimation of competence, or objections for

estimation of a level of satisfaction.

Subjective expert opinion(s) may be obtained by means of manual evaluation of a set of dialogues. For this, a fixed scale is taken and each dialogue is evaluated in the framework of this scale. Usually symmetric normalized scale $[-1,1]$ or positive normalized scale $[0,1]$ is considered.

In order to construct an empirical formula we use the Inductive Method of Model Self-Organization (IMMSO) proposed by Ivakhnenko [22]. This method allows to select the best formula from a given class using the training and the control sets of examples.

For definiteness, in this paper we consider only client's politeness. And it should be emphasized that we have no aim to find the best way for numerical estimation of politeness. Our goal is only to demonstrate how one may transform the lexical-grammatical properties of a text and the subjective expert opinion to these numerical estimations.

The paper is organized as follows. Section 6.2 describes the linguistic factors that should be taken into account in the formula to be constructed. In Section 6.3, we show how to apply Ivakhnenko method to the problem of politeness estimation. Section 6.4 contains the results of experiments. Conclusions and future work are drawn in Section 6.5.

6.2 Models for parameter estimation

6.2.1 Numerical indicators

The model to be constructed represents a numerical expression, which depends on various indicators of politeness of a given text and determines a certain level of politeness. This level is measured by a value between 0 and 1, where 0 corresponds to a regular politeness, and 1 corresponds to the highest level of politeness. We do not consider any indicators of impoliteness, although in some cases it should be done.

In this paper we take into account the following 3 factors of politeness: the first greeting (**g**), polite words (**w**) and polite grammar forms (**v**). As examples of polite words such well-known expressions as "please", "thank you", "excuse me", etc. can be mentioned. We considered verbs in a subjunctive mood as the only polite grammar forms, e.g. "could you", "I would", etc.

We take into account the following two circumstances:

(i) The level of politeness does not depend on the length of the dialogue. It leads to the necessity to normalize a number of polite expressions and polite grammar forms by the length of dialogue. The dialogue's length here is the number of client's phrases.

(ii) The level of politeness depends on the number of polite words and polite grammar forms non-linearly: the greater number of polite words and grammar forms occur in a text the less contribution new polite words and grammar forms give. It

leads to the necessity to use any suppressed functions as the logarithm or the square root, etc.

Therefore, we consider the following numerical indicators of politeness:

$$\mathbf{g} = \{0, 1\}, \quad \mathbf{w} = \ln(1 + N_w/L), \quad \mathbf{v} = \ln(1 + N_v/L), \quad (6.1)$$

where N_w, N_v are a number of polite words and polite grammar forms respectively and L is a length of a dialogue.

It is evident that: a) $\mathbf{w} = \mathbf{v} = 0$, if polite words and polite grammar forms do not appear; b) $\mathbf{w} = \mathbf{v} = \ln(2)$, if polite words and polite grammar forms occur in every phrase. All these relations are natural and easy to understand.

6.2.2 Example

In this section, we demonstrate how the mentioned indicators are manifested and evaluated. Table 6.1 shows the example of dialogue (the records are translated from Spanish into English). Here US stands for a user and DI for a directory inquiry service.

Table 6.1: Example of a real dialogue between passengers and directory inquiries

US: Good evening. <i>Could you</i> tell me the schedule of trains to Zaragoza for tomorrow?	DI: I will see, one moment. The next train leaves at 5-30
DI: For tomorrow morning?	US: 5-30
US: Yes	DI: hmm, hmm ; SIMULTANEOUSLY ;
DI: There is one train at 7-30 and another at 8-30	US: Well, and how much time does it take to arrive?
US: And later?	DI: 3 hours and a half
DI: At 10-30	US: For all of them?
US: And till the noon?	DI: Yes
DI: At 12	US: Well, <i>could you</i> tell me the price?
US: <i>Could you</i> tell me the schedule till 4 p.m. more or less?	DI: 3800 pesetas for a seat in the second class
DI: At 1-00 and at 3-30	US: Well, and what about a return ticket?
US: 1-00 and 3-30	DI: The return ticket has a 20% of discount
DI: hmm, hmm ; SIMULTANEOUSLY ;	US: Well, so, it is a little bit more than 6 thousands, no?
US: And the next one?	DI: Yes
	US: Well, <i>thank you very much</i>
	DI: Don't mention it, good bye

Table 6.2 shows the results of parameterization of this dialogue and its manual estimation by a user. Here the number of polite words is equal to 2 because the

passenger utilized the polite form of a pronoun “you” that has no analogue in English.

Table 6.2: Example of a real dialogue between passengers and directory inquiries

First greating \mathbf{g}	Number of polite words N_w	Number of polite grammar forms N_v	Indicator \mathbf{g}	Indicator \mathbf{w}	Indicator \mathbf{v}	Estimation
Yes	2	2	1	0.13	0.13	1

In our work, all the factors \mathbf{g} , \mathbf{w} , \mathbf{v} are detected by means of the NooJ resource [34] (previously, for the same purpose we used morphological analyzers described in [16]). The NooJ is a linguistic tool to locate morphological, lexical and syntactic patterns used for raw texts processing. The results of the NooJ analysis were fixed in a file for further processing by Ivakhnenko method.

6.2.3 Numerical models

Taking into account the three factors described above the following series of polynomial models can be suggested for automatic evaluation of the level of politeness:

$$\begin{aligned}
 \text{Model1 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 \\
 \text{Model2 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= C_0 \mathbf{g} \\
 \text{Model3 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} \\
 \text{Model4 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} + B_{10} \mathbf{w} + B_{01} \mathbf{v} \\
 \text{Model5 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} + B_{10} \mathbf{w} + B_{01} \mathbf{v} + B_{11} \mathbf{vw} \\
 \text{Model6 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} + B_{20} \mathbf{w}^2 + B_{02} \mathbf{v}^2 \\
 \text{Model7 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} + B_{11} \mathbf{vw} + B_{20} \mathbf{w}^2 + B_{02} \mathbf{v}^2 \\
 \text{Model8 : } F(\mathbf{g}, \mathbf{w}, \mathbf{v}) &= A_0 + C_0 \mathbf{g} + B_{10} \mathbf{w} + B_{01} \mathbf{v} + B_{11} \mathbf{vw} + B_{20} \mathbf{w}^2 + B_{02} \mathbf{v}^2 \\
 &\text{etc.}
 \end{aligned} \tag{6.2}$$

Here: A_0 , C_0 , B_{ij} are undefined coefficients. As it may be observed all these models are the polynomials with respect to the factors \mathbf{w} and \mathbf{v} . Such a representation is rather general for various functions $\psi(\mathbf{w}, \mathbf{v})$ and this is a reason of its application. Of course, one can suggest another types of models.

6.3 Application of IMMSO

There are two variants of the Ivakhnenko method:

- (i) Combinatorial Method;
- (ii) Method of Grouped Arguments.

In the first case, the sequence of models is considered step-by-step, while in the second one, the models are filtered [23]. In this work, we use only the first method and consequently consider all 8 models (6.2) presented in Section 6.2.3.

Parameters of the concrete model are determined by means of the least square method. For that, we fix one of the models (6.2) and construct the system of lineal equations for a given set of dialogues:

$$F(\mathbf{g}_i, \mathbf{w}_i, \mathbf{v}_i) = P_i, \quad i = 1, \dots, N \quad (6.3)$$

where $\mathbf{g}_i, \mathbf{w}_i, \mathbf{v}_i$ are the factors, P_i are the manual estimations of dialogue, N is the number of dialogues. For example, the dialogue described in Tables 6.1,6.2 forms the following equation for the 4th model : $A_0 + C_0 + 0.13B_{10} + 0.13B_{01} = 1$

The system (6.3) is a system of lineal equations with respect to undefined coefficients. It can be solved by the least square method. It should be taken into account that the number of equations must be several times greater than the number of parameters to be determined. It allows to filter a noise in the data. By 'noise' we mean, first of all, fuzzy estimations of politeness.

According to IMMSO methodology for the series of models starting with the first model from (6.2), some external criterion is calculated and checked whether it achieves an optimal point. Depending on the problem different forms of this criterion can be proposed [23]. In our case, we use the criterion of regularity. It consists in the following:

1. model parameters (coefficients A_0, C_0 , etc.) are determined on the training data set;
2. this model is applied to control data set and 'model' politeness is calculated;
3. the relative difference between the model politeness and the manual politeness of an expert is estimated.

All these actions can be reflected by the following formula:

$$K_r = \frac{\sqrt{\sum_N (P_i(T) - P_i)^2}}{\sqrt{\sum_N (P_i)^2}} \quad (6.4)$$

where $P_i(T)$ are the 'model' estimations of politeness on the control data set, that is the left part of the equations (6.3), P_i are the manual estimations of dialogues from the control data set, N is the number of dialogues in control data set. It should emphasize that the model parameters are determined on the training data set.

6.4 Experiments

The data we used in our experiments represent a corpus of 100 person-to-person dialogues of Spanish railway information service. Some characteristics of the corpus (length of talking, volume of lexis) are described in [4]. From the mentioned corpus we took randomly $N = 15$ dialogues for training data set and $N = 15$ dialogues for control data set. The level of politeness was estimated manually in the framework of scale $[0, 1]$ with the step 0.25. Table 6.3 represents a part of data used for the experiments.

Table 6.3: Example of data used in the experiments

\mathbf{g}	\mathbf{w}	\mathbf{v}	\mathbf{w}^2	\mathbf{wv}	\mathbf{v}^2	Manual estimation
1	0.134	0.194	0.0178	0.0259	0.0377	1
0	0.111	0.057	0.0124	0.0064	0.0033	0.75
1	0.000	0.074	0.0000	0.0000	0.0055	0.25
1	0.000	0.031	0.0000	0.0000	0.0009	0
1	0.000	0.118	0.0000	0.0000	0.0139	0.75
1	0.043	0.043	0.0018	0.0018	0.0018	0.5
1	0.000	0.000	0.0000	0.0000	0.0000	0.25
1	0.043	0.083	0.0018	0.0035	0.0070	0.5
0	0.000	0.074	0.0000	0.0000	0.0055	0
1	0.134	0.069	0.0178	0.0092	0.0048	1

We tested all 8 models (6.2) and calculated the criterion of regularity (6.4). The results are presented in Table 6.4.

Table 6.4: Values of the regularity criterion for polynomial models of different complexity

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
0.505	0.567	0.507	0.253	0.272	0.881	1.875	0.881

It may be observed that the criterion of regularity achieves its minimum on the lineal model (Model 4). The fact that the most appropriate model reflects only trend could be explained by imperfectness of a given class of models and/or a high level of noise. Joining together all 30 examples we determined the final formula as:

$$F(\mathbf{g}, \mathbf{w}, \mathbf{v}) = -0.14 + 0.28\mathbf{g} + 3.59\mathbf{w} + 3.67\mathbf{v} \quad (6.5)$$

This formula provides 24% of relative mean square root error.

In order to evaluate the sensibility of the obtained results to the volume of data the same calculations were accomplished on the basis of 10 dialogues in the training set and 10 in the control one. We considered only first 4 models, since more complex

models need more data. The results presented in Table 6.5 show that the dependence on the volume is insignificant with respect to the behaviour of external criterion.

Table 6.5: Criterion of regularity for the reduced data set

Model 1	Model 2	Model 3	Model 4
0.497	0.503	0.502	0.319

6.5 Conclusions

In this chapter, we demonstrate another application of IMMSO to the task of dialogue processing, namely, we suggest a simple methodology for an automatic estimation of various 'fuzzy' dialogue characteristics, which have a large level of subjectivity. The constructed formula for politeness estimation correctly reflects the contribution of selected factors of politeness: all factors have positive coefficients. The obtained error is comparative with the step of the manual dialogue estimation.

In the future, we intend to consider more complex empirical models for estimation of politeness, satisfaction, culture and competence.

Chapter 7

Conclusions and future work

Ontologies play one of the principle roles in a future development of Semantic Web technologies. In our work, we deal with some aspects of ontology learning. We concentrate our attention on two problems of ontology construction: term's recognition for special domains of higher difficulty and exploration of granularity of ontological concepts.

Together with resolving the above problems we explore stability of IMMSO and also demonstrate its use for a task of dialogue processing.

We can resume our main contributions as follows:

1. A novel HMM-based approach for biomedical NER.
2. Comparing performance of different ML methods under the same conditions in a biomedical NER task.
3. A formal definition of granularity of domain terms.
4. A method of revealing granularity levels of domain terminology.
5. Elaborating stability of IMMSO for different parameters of initial data.
6. Constructing an empirical formula for estimating client's characteristics in dialogue processing.

Our future work include a further investigation of ontology properties that can be and are needed to be optimized. As far as the characteristic of granularity is concerned we plan to continue our investigation taking into account not only term's granularity but also granularity of ontological relationships.

Briefly, we are going to accomplish the following tasks:

1. Revealing granularity levels of ontological relationships.
2. Exploiting Latent Semantic Analysis (LSA) for discovering topics of distinct granularity in a given corpus.
3. Mapping types of relationships between strong correlated terms.

Bibliography

- [1] M. Alexandrov, X. Blanco, N. Ponomareva, and P. Rosso. Constructing empirical models for automatic dialog parameterization. In *Proceedings of Text, Speech, Dialog (TSD-07)*, LNCS. Springer Verlag, 2007.
- [2] M. Alexandrov, E. Sanchis, and P. Rosso. Cluster analysis of railway directory inquire dialogs. In *Proceedings of the TSD'05*, pages 385–392, 2005.
- [3] G. Bisson, C. Nedellec, and L. Canamero. Designing clustering methods for ontology building - the mo'k workbench. In *Proceedings of the ECAI Ontology Learning Workshop*, 2000.
- [4] A. Bonafonte. Desarrollo de un sistema de dialogo oral en dominios restringidos. In *I Jornadas en Tecnologia de Habla*, 2000. (in Spanish).
- [5] J. Brank, M. Grobelnik, and D. Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, Ljubljana, Slovenia, 2005.
- [6] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data driven ontology evaluation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal, 2004.
- [7] A. Burton-Jones, V. Storey, V. Sugumaran, and P. Ahluwalia. A semiotic metrics suite for assessing the quality of ontologies. In *Data and Knowledge Engineering*, 2005.
- [8] S. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL) Data and Knowledge Engineering*, 1999.
- [9] E. Charniak and M. Berland. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

-
- [10] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2004.
- [11] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.
- [12] K. B. Cohen and L. Hunter. *Natural Language Processing and Systems Biology*. Springer Verlag, 2004.
- [13] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- [14] Hartmann et al. Methods for ontology evaluation. *Knowledge Web Deliverable D1.2.3*, 2005.
- [15] D. Faure and C. Nedellec. A corpus-based conceptual clustering method for verb frames and ontology. In P. (Ed.) Velardi, editor, *Proceedings of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, 1998.
- [16] A. Gelbukh and G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. *Springer, LNCS*, N 2588:215–220, 2003.
- [17] T.R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.
- [18] N. Guarino and C. Welty. *Ontology learning*, pages 151–172. 2002.
- [19] Z. Harris. *Mathematical Structures of Language*. Wiley, 1968.
- [20] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 1992.
- [21] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1990.
- [22] A. Ivahnenko. *Manual on typical algorithms of modeling*. Tehnika Publ., 1980. (in Russian).
- [23] A. Ivahnenko. *Inductive method of model self-organization of complex systems*. Tehnika Publ., 1982. (in Russian).

-
- [24] J. D. Kim, T. Ohta, Y. Tsuruoka, and Y. Tateisi. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the Int. Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 70–75, 2004.
- [25] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [26] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*, pages 282–289, 2001.
- [27] A. Lozano-Tello, A. Gomez-Perez, and E. Sosa. Selection of ontologies for the semantic web. In *Proceedings of ICWE-2003*, volume 2722 of *LNCS*, pages 413–416. Springer Verlag, 2003.
- [28] A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of CIKM-2002*, volume 2473 of *LNAI*, pages 439–448. Springer Verlag, 2002.
- [29] P. Makagonov and M. Alexandrov. Constructing empirical formulas for testing word similarity by the inductive method of model self-organization. In *Proceedings of Advances in Natural Language Processing*, volume 2389 of *LNAI*. Springer Verlag, 2002.
- [30] A. McCallum. Efficiently inducing features of conditional random fields. In *In Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence (UAI-2003)*, 2003.
- [31] D. Mladenic and M. Grobelnik. Feature selection for classification based on text hierarchy. In *Working Notes of Learning from Text and the Web, Conf. Automated Learning and Discovery (CONALD-98)*, Carnegie Mellon Univ., Pittsburgh, US, 1998.
- [32] A. Molina and F. Pla. Shallow parsing using specialized hmms. *JMLR Special Issue on Machine Learning approaches to Shallow Pasing*, 2002.
- [33] A. Montejo-Rez, L. A. Urea-Lpez, and R. Steinberger. Categorization using bibliographic records: beyond document content. *Procesamiento del Lenguaje Natural*, 35(1):119–126, 2005.
- [34] NooJ. <http://www.nooj4nlp.net>.
- [35] A. Orme, H. Yao, and L. Etzkorn. Indicating ontology data quality, stability, and completeness throughout ontology evolution. *Journal of Software Maintenance and Evolution: Research and Practice*, 19:49–75, 2007.

-
- [36] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1993.
- [37] M. Poesio, T. Ishikawa, S. S. im Walde, and R. Viera. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*, 2002.
- [38] R. Porzel and R. Malaka. A task-based approach for ontology evaluation. In *Proceedings of ECAI-2004 Workshop Ontology Learning and Population*, 2004.
- [39] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257–285, 1998.
- [40] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *In Advances in Neural Information Processing (NIPS17)*, 2004.
- [41] B. Settles. Biomedical named entity recognition using conditional random fields and novel feature sets. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 104–107, 2004.
- [42] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *In Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03)*, 2003.
- [43] C.E. Shannon. A mathematical theory of communication. *Bell System and Technical Journal*, 27:379–423, 623–656, 1948.
- [44] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.
- [45] P. Spyns. *EvaLexon: Assessing triples mined from texts*. STAR Lab, Brussels, Belgium, 2005. Technical Report 09.
- [46] M. Uschold and M. Gruninger. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.
- [47] S.K.M. Wong and Y.Y. Yao. An information-theoretic measure of term specificity. *Journal of the American Society for Information Science*, 43(1):54–61, 1992.

-
- [48] Y. Yang and J.O. Pedesen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420, 1997.
- [49] J. Zhang, D. Shen, G. Zhou, S. Jian, and C. L. Tan. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6), 2004.
- [50] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 96–99, 2004.