AN APPROACH TO NEW ONTOLOGIES DEVELOPMENT: MAIN IDEAS AND SIMULATION RESULTS

B. Dobrov, N. Loukachevitch, O. Nevzorova

Abstract: In the paper we consider the technology of new domain's ontologies development. We discuss main principles of ontology development, automatic methods of terms extraction from the domain texts and types of ontology relations.

Keywords: Ontology, thesaurus, automatic term extraction, ontology relations.

Introduction

The present article is dedicated to the technology of creation of the so-called linguistic ontologies, i.e. ontologies, concepts in which are generally based on the semantics of the domain terms. Such ontologies are usually used for the automatic processing of texts in a natural language.

The technology has been developed in the process of creation of large and extra large ontologies and thesauri for various domains and their actual usage in multiple applications of automatic text processing.

Among such works are the following:

thesaurus of the social and political life (28 thousand concepts, 67 thousand terms, 100 thousand conceptual relations), - Sociopolitical thesaurus [Loukachevitch, 2002], which is a search means in the University information system RUSSIA (www.cir.ru) and is used in such applications of automatic text processing as conceptual indexing, automatic text categorization, automatic text summarization;

the Russian language thesaurus RuThes (43 thousand concepts, 100 thousand words and expressions, 166 thousand conceptual relations) [Loukachevitch N., 2002];

thesaurus for the domain «Elections» - is included into the Sociopolitical thesaurus;

thesaurus for the domain «Economic statistics»;

ontology for the domain «Software functionality» for decision making support during software testing;

Avia-Ontology for the domain, describing behavior of an operator (air crew) and board equipment in various flight operations (1200 concepts, 3400 terms). Aim of the Avia-Ontology development is the analysis of the completeness of documents, describing the logic of work in typical flight regimes [Nevzorova, 2001].

Avia-Ontology is currently being developed and will be used as a basic source of examples for the article [Dobrov, 2002].

It should be highlighted that the peculiarity of the proposed technology and the existing experience is namely the activity of a knowledge engineer, who at the beginning of work has a very superficial idea of the conceptual structure of a domain and its terminology.

1. The formation of text collection

One of the essential conditions of the successful development of a linguistic ontology is preliminary creation of an electronic text collection with reference to a domain. The collection may be of various genres and may include textbooks, scientific articles, technical data, mass media works and so on.

During the development of the Sociopolitical thesaurus we used the text collection of the University Information system RUSSIA, comprised of over 700 thousand documents: official documents, laws, scientific works on social studies, newspaper releases.

During the development of the Avia-Ontology a great effort had to be made in order to form a sufficient collection of electronic documents on the given domain. To a great degree scanning of the printed matter and search of relevant materials in Internet had to be done. As a result an electronic collection with the size over 100Mb was formed.

2. Automatic terms extraction from the domain texts

2.1. Terms extraction based on syntactic information

Having formed an electronic collection of domain texts it is necessary to obtain its "terminological portrait", for which the procedures of automatic extraction of terms, essential for the given domain, are employed. For the Russian language terminology the syntactic structure of over 90 percent of domain terms [Loukachevitch, 2002] refers to one of the following constructions:

single nouns, adjectives, and words not described in the morphological dictionary – usually abbreviations;

noun phrases (NP): noun + noun in the genitive case;

NP: adjective + noun

NP: adjective + adjective + noun

NP: noun + adjective + noun in the genitive case

Such types of constructions are collected on the basis of preliminary morphological processing of texts. In this process noun and adjective agreement was checked, after which the bulk of such noun phrases are syntactically correct groups. Decreasing frequency-ordered lists of such words and word expressions present important data for the formation of an idea about the domain.

It should be mentioned that the list includes many non-terminological linguistic expressions.

First of all, domain texts contain a great number of words of general meaning, e.g. *possibility, means, condition, type, point* and others.

The larger a domain and a text collection are, the bigger problem is to exclude multiword constructions of general meaning, such as *task solution, further development, uniform system, beginning of the year, present time*. During the development of Sociopolitical thesaurus we made use of special lexical filters, organized as a specialized vocabulary to exclude such general expressions from consideration [Loukachevitch, 2002]. Such filters are undoubtedly useful for the creation of extra large ontologies, however, they also depend on a domain. So, for instance, the word *argument* may be considered as a non-terminological one in the sociopolitical domain, however, *argument* is an important term in the domain of software production.

Besides, in a domain there may be longer terms or terms of a different syntactic construction, like those containing prepositions.

It should however be highlighted that the simplicity of the given algorithm of term extraction is an important factor of its use for the analysis of the domain structure.

2.2. Multiword terms extraction on the basis of the text structure

In order to extract longer terms and/or terms containing prepositions another method is used.

Many algorithms of multiword term extraction use the supposition that words, comprising a term, frequently occur together. [4, 5]. To find terms of a more complex syntactic structure that described earlier, we use our own variant of algorithms of such type.

If the text author use a certain term as a separate unit of narration, then exactly in this text words of the term will occur alongside more frequently than spaced out.

To reveal this during the text processing for every word (noun, adjective) an immediate neighbor word and neighbor-words in the text window of a given size are stored. A table of immediate neighbor words and a neighbor table in the text window are created and the frequency of word pairs occurrence is calculated.

Further it is expected that if a pair of words occurs as immediate neighbors in over half of cases of their occurrence in the same text window, this proves that this pair in aggregate serves as a reference point in the text, i.e. represents a term or a fragment of a term.

In this case the word pair is glued together to form a common terminological unit and the tables are recalculated as if this unit has been known from the very beginning before the text processing. This step gives the possibility to the further development of the term, thus forming units of length 3 and longer. Examples of terms obtained in this way in the aviation domain are as follows: *break off the attack*, *position of tactical advantage, flying in pairs* and so on.

In the domain “Elections” such terms as *member of election committee with right to deliberative vote, executive body of local government, local governor’s elections* were obtained.

When being reviewed, the terms obtained are ordered not by the frequency of their occurrence, but by the number of texts in which they occurred. It is supposed that a special attention should be paid to those word-combinations which remained stable in over two texts. As a result of the experiment on the text collection of 50Mb newspaper releases 1346 such word-combinations were obtained. 80% of them were qualified as terms. The comparison was drawn with the Sociopolitical thesaurus and there were singled out approximately 30 important terms, not present in it at that moment, such as *volume of output, economy in transition, capital flow.*

The algorithm drawback is that it practically does not extract terms from short texts - a term has to be used in a text at least twice, better three times. However this drawback is insignificant when dealing with extra large collections, such as the collection of UIS RUSSIA. Thus we suggest to process the whole collection using the abovementioned algorithm to single out terms not present in the Sociopolitical thesaurus and to further monitor the appearance of new terms in the sociopolitical domain.

We consider useful to apply both described algorithms to work with small domains.

3. Using general linguistic resource as a base for the development of applied ontology

The terminology of any domain contains both specific terms, used only in the given domain or in the range of similar domains, and rather commonly known terms. In the given domain the examples of such commonly known terms are *pilot, airplane, pursuit plane, weapon, attack* and many others. This allows us not to begin the domain model development from point zero, but to use the knowledge described in more general linguistic and terminological resources.

As such a source we used the Thesaurus of the Russian language RuThes [Dobrov, 2002]. The Thesaurus represents a hierarchical network of concepts, each of which has a number of text variants (linguistic expression means) and conceptual relations with other concepts of the thesaurus. The resource volume nowadays makes up 100 thousand words and word expressions, confined to 43 thousand concepts. Over 166 thousand relations were manually determined between the concepts. Over 1200000 relations between concepts of the Thesaurus were established on the basis of the transitivity and inheritance properties.

The RuThes Thesaurus contains two big parts. RuThes is comprised of the Sociopolitical thesaurus, including terminology of economic, political, military, social, scientific and other spheres (64 thousand words and terms). The zone of words and word expressions of the thesaurus, designating actions, situations, objects, which can occur in any subject area texts, and thus not included in the Sociopolitical thesaurus, is called general lexicon (33 thousand words and word expressions).

The presence of a big generally valid linguistic resource makes possible to compare the given resource with the domain texts, to single out the knowledge types described in the thesaurus (concepts, synonyms, relations between concepts), to transfer them to a special working domain as a basis for the creation of a domain ontology. Lists of gathered words and word-combinations in the domain were compared to the terms of the Sociopolitical thesaurus. If the comparison of the next word-combination was successful, a corresponding concept from the thesaurus together with all the terms that express it in the text, was copied into the domain model. During next step all the relations of the Sociopolitical thesaurus between the copied concepts were copied.

Besides, the closing of relations was performed: if concept B is superior compared to concept A, concept C is superior to concept B, while concepts A and C were copied into a domain, then concept B is also copied into the domain model together with its terms variants and relevant relations.

The comparison with the general lexicon zone was controlled manually, as part of such words preserved its general meaning, for instance *necessity, conditions*, etc. and , thus, do not need to be reflected in the domain model, other words like *chandelle, cover, escort* are important concepts of the domain, which therefore have to be reflected in the model. In the second case the corresponding concepts were also copied together with their terms and relations, just as during the comparison with the Sociopolitical thesaurus terms.

Undoubtedly, the transferred concepts and relations require a thorough additional test for adjusting the setup to a certain domain. For instance, as a result of transfer among the terms we can come across a synonymic variant, which is improbable in a given domain, for example, the word «rotorcraft» as a synonym of the word «helicopter» is unlikely to be found in the professional technical area.

In the present version of the Avia-Ontology about one-third (1100) terms were transferred from the thesaurus RuThes together with synonymic relations and relations between concepts, which provided fast development of the Avia-Ontology and considerable reduction of the new resource development time.

4. The main procedure of the ontology construction

Having exhausted the available linguistic resources, we begin analyzing the available texts and terms lists for the further ontology development.

The decision making in this process consists of the following steps:

on the basis of the available text material we look for a word or a word-combination, designating an important concept (how to determine the concept «importance», we will consider in the following sections).

Entering a new concept we must provide it with an understandable and unambiguous name. If possible it is better to enter the concept names which are unambiguous even outside the current domain. For instance, when entering the concept "airplane wing" it is possible to name it "*wing"* – in the given domain this word is unambiguous, but it is better to give it a clearer name, which will not lose its unambiguity with the domain expansion or after including this ontology into a larger ontology;

when entering a concept at least one relation of this concept with the other ontology concepts is entered. On the one hand, it is supposed that if it is rather complicated to set a relation for the entered concept, then it is too early to enter such a concept and additional analysis is required. On the other hand there is no need (and as a rule it is impossible) to describe all the necessary relations of a new concept at once. Practice proves that by entering further new concepts the initial position may become clearer – in order to enter something new, it is often necessary to correct the inaccuracies and distortions of the old. Properly speaking, inclusion of new concepts helps to reveal the problems of the existing description;

- and, at last, a concept must be supplied with a list of words and multiword expressions, which can be used to refer to the entered concept in texts. As such text entries single words (nouns, adjectives, verbs), noun and verb phrases can be included. We suppose that a multiword linguistic expression must be used in a text as an inseparable construction. A text entry may be ambiguous (may have another meaning), then it must be specially marked. Besides, a sequence of normalized forms of all constituents of a multiword expression must be entered (masculine gender, nominative case, singular), which will be used for term recognition in texts.

5. Ontology concepts selection

5.1 Single word-based concepts

The single words, occurring in domain texts, have two subgroups, on which it is easy to decide whether to include or not include the corresponding concepts into an ontology.

One of such groups - evident terms of a given domain, for example, *aircraft, flaps,* and corresponding concepts must be included in an ontology. Another group – evident words of general meaning, such as *necessity, possibility, creation, etc.,* for which inclusion into an ontology is not necessary.

It is more problematic to make a decision on the other groups.

One of such groups - terms of a given domain that appear on the basis of general lexicon words like *turn, reversal, dive*. Narrowing or other changes of basic meanings are characteristics of such words. So, in the aviation domain *turns* refer to the airplanes. Besides, to help in distinguishing such terms serves the fact that in the terminological word-combinations lists there is a considerable amount of different word-combinations with the inclusion of this word, which should also enter the ontology: *turn, combat turn, turn radius.*

Another «complex group» of words are words that clearly refer to the vocabulary of a general meaning, but ontology includes a certain number of concepts, based on the terms with this word: break off – *break off combat, break off attack*. The questions arise if corresponding generalizing concepts should enter an ontology. Two aspects should be considered. On the one hand, the appearance of such a generalizing

concept provides an additional structure to the ontology, which is a positive factor. However, on the other hand, if a word is very ambiguous in the framework of a given domain, and abstract, then it may cause serious problems in the lexical ambiguity resolution, and this corresponding concept should not enter the ontology.
The last group of words, which requires a special effort to decide if a corresponding concept should be included into an ontology or not, are words, that are on the borderline between domains or those that were relatively accidentally used in a text collection, for instance, *aircraft construction* (whether refers to the subject area or not), *adapter, heat insulation* and so on.

5.2. Multiword word-combinations-based concepts

Any domain text contains a great number of various multiword linguistic expressions. Selection of such expressions for the inclusion into an ontology is a serious problem. Existing terminological lists of a domain usually embrace only a small part of those term-like expressions that are met in texts. Experts may not also have a certain opinion on the bulk of such expressions. Therefore it is necessary to have a total of principles to decide which specific factors are to be taken into account for including concepts based on multiword expressions into an ontology.
We should highlight the main principle at this point. If such a concept is included into an ontology, it should happen not so much because the corresponding word-combinations refer or not to the vague category of the given domain terms, as which new information the appearance of this concept in the ontology gives. Thus, a new concept in an ontology is the application point of additional information which can be used in automatic text processing.
Such information may be divided into several types.

5.2.1. Existing and important

Any domain has a small number of main points which are extremely important in the given domain. Terms and other linguistic expressions that correspond them are highly frequent in the subject area texts. Such main points (concepts, single objects) must be reflected in the ontology. So, working in the domain «Elections» it is essential to have in the ontology a conceptual unit *CENTRAL ELECTION COMMITTEE OF THE RUSSIAN FEDERATION.*
If concepts entered into the ontology have a fixed and small number of narrower concepts, then they have to be reflected in the ontology. So, for the domain «Elections» types of elections are reflected, in the Sociopolitical thesaurus – types of budget, in the subject area of military aviation - bombing flight regimes *(dive bombing*).
Another important type of information is that two concepts have a common subtype. For instance, concepts *DEFENSIVE MANEUVER* and COMBAT TURN have a subtype *DEFENSIVE TURN*, concepts *FINE* (*PENALTY)* and *ADMINISTRRATIVE PUNISHMENT*– subtype *ADMINISTRATIVE FINE.*

5.2.2. Multiword expression has «interesting» synonyms

A concept can unite the total of various text expressions with the same meaning (derivatives (*to take off, take off*) are also considered as text entries of the same concept). So, the revealing of synonymic expressions or derivatives often leads to the introduction of a new concept for the fixation of the synonymy found. The variety of textual expressions in this case often points at the importance of a corresponding concept.
After the concept is set up, special effort is applied to find other ways of referring to the same concept in texts, i.e. the synonymic raw of the concept is maximally filled. These variants may seem obvious for a person, and their entry may seem tiresome, but as practice proves, during automatic processing of various texts direct matching is better that any logical inference. It is often supposed that this or that variant exists, and then its actual occurrence is checked by Internet. For instance, if a new concept is entered on the basis of the term *horizontal flight bombing,* then instantly existence of the synonym *horizontal bombing,* which actually exists, is checked. If afterwards a concept for the term *pitch-up bombing* is entered, then, naturally, the existence of the term *pitching bombing* is checked; such term was not found in Internet. Let us give example of synonymic row:

ENGINE POWER INCREASE
Increase engine power
Increase engine thrust
Increase thrust
Engine acceleration

5.2.3. Relations that do not follow from the structure of a multiword expression

The principle used to evaluate the necessity of entering a concept into many thesauri and ontologies is that a multiword term has relations that do not follow from its structure.
Examples of such relations are:
accelerated turn – loss of speed,
attack evasion – defensive maneuver,
superiority in energy – tactically advantageous position.
To fix this relation it is necessary to introduce the corresponding concepts.

5.2.4. Completion of ontology levels

An important principle of ontology completion is the "closing" principle, which has two subtypes.
In the first place, if a new concept, introduced by any reason has created a new inferior ontology level, then it has to be completed by other essential concepts of the same level. For instance, if *MISSILE LAUNCH* concept is entered as an inferior one for the concept *USE OF WEAPON,* then it is necessary to enter, for example, the concept *CANNON FIRING*, as the second most important type of using weapons in the given domain.
This principle is at the same time limiting: if we decide to enter a new-level concept, we must evaluate the consequences of such a step: how many concepts of the same level we are going to enter; if the number of potential concepts of this level is too big, then the entry limiting principles should be determined at once. So, for instance, in the *Sociopolitical thesaurus* there can be a lot of concepts inferior to the concept *GOODS FOR CHILDREN*. The appearance in the inferior row of concepts *CLOTHES FOR CHILDREN, SHOES FOR CHILDREN, TOYS FOR CHILDREN* is additionally justified by the existence and inclusion into the thesaurus of certain types of these goods, having single lexemes as text entries.
On the other hand, an opposite situation may emerge: several concepts sharing common features are found, it is necessary to find a common concept. For instance, on the basis of subject area texts analysis concepts *FLAPS* and *SLATS* are introduced, common features of which is that both are located on the wings and that they serve to control a flight. Extra attention is paid to searching for a generalization, which is found - *FLIGHT CONTROL SURFACES.*
Another example of generalizing two concepts that were entered: *AFTERBURNER LIGHTING* and *DECREASE THRUST* lead to the entry of the concept *THRUST CONTROL*.

5.2.5 Single words are ambiguous and word-combination is unambiguous

An important factor which helps to determine the entry of a new concept is the presence of ambiguous words inside an unambiguous multiword expression.
So, an ambiguous term *press* is important for the sociopolitical area, and to support the ambiguity resolution process we introduced such concepts as *LOCAL PRESS, CENTRAL PRESS, ILLEGAL PRESS* into the Sociopolitical thesaurus.
Working in the same broad sociopolitical area we may hesitate if the introduction of concepts *SHORT FILM–FEATURE FILM* is necessary, but as soon as it becomes known that *SHORT FILM* has the synonym *short subject*, the corresponding concepts are included at once.
We should emphasize that all the abovementioned does not mean that for every word-combination, consisting of ambiguous words, a corresponding concept is created; the described principle only helps to make the decision in such cases when we are almost ready to introduce a concept.

6. Ontology relations

The most important part of the prevailing number of ontologies is the total of relations between the concepts. This set of relations largely depends on the domain and on the task for solving which ontology is meant. We suggest to begin the construction of an ontology on a minimal set of relations and to determine the domain structure according to this set. Such a minimal set of relations does not depend on the type of a domain, on the type of a problem solved, as it is based on the fundamental properties of concepts. In the first place for a given concept we determine such concepts on which depend its existence or existence of the given concept examples, i.e. determining the so-called relations of ontological dependence, which are studied in the framework of the philosophical discipline «formal ontology»

The main instruments of entities analysis within Formal Ontology [Smith, 1998] are the following:

the theory of identity, integrity. The main problems of this type of analysis: what does the fact that two entitles are one and the same thing means, how can an entity change and preserve its identity, what properties are essential for preserving one's identity, etc.

the theory of part and the whole (mereology, mereotopology). The main problems here are the following: what is considered as a whole, what makes an entity a whole, what is the connection of parts in the whole, what properties such a connection relation has, how is the whole separated from the «background», what are boundaries and so on.

the dependence theory [Guarino, 1998].

The main question of the dependence theory is if an entity can exist by itself or it supposes the existence of something else:

whether the existence of an entity supposes the existence of something else (rigid dependence), for instance, *boiling* is impossible without the existence of a certain volume of liquid which boils;

whether existence of examples of a certain class (generic dependence) is supposed, like, the appearance of the concept *garage* is impossible without the existing concept *motor vehicle,* though a certain garage may appear without any reference to a certain motor vehicle;

- when the existence of an entity in moment T presumes the existence of another entity in moment T1 before T (historical dependence), so, for instance, straw historically depends on threshing, as straw can not appear without a preliminary threshing process, altogether this work comes to an end, while straw continues its existence for a long time.

Examples of conceptual dependence relations in the Avia-Ontology are as follows:

ALTIMETER depends on *FLIGHT ALTITUDE* (generic dependence),

TANKER AIRCRAFT depends on AIRCRAFT FUEL (generic dependence),

AIR PATROL depends on *FIGHTER AIRCRAFT (*rigid dependence*)*

Thus, for each ontology we suggest to develop a sort of an "initial" ontology, in which non-taxonomic relations are relations of the conceptual dependence. Such ontology can serve as a basis for explication of the domain structure and determination of a new set of relations, necessary for solving the main problem. In this case the conceptual dependence relations are so important for any domain, that there is no need to delete them, it is only necessary to re-name them in the newly introduced relations system.

A specific set of relations, which is used by us now besides taxonomic relations (BROADER-NARROWER relations), is the following:

PART– WHOLE – is used to describe the traditional parts, participants of situations, properties. Here the conceptual dependence of concept-part on the concept-whole is required;

unsymmetrical associations ASC1-ASC2 – are used for the rest of conceptual dependence relations;

symmetric association is used for concepts, similar by meaning.

Thus, two types of relations in the relations set employed by us are significantly bound with the concept of ontological dependence. Relations of these types occupy approximately half of all relations in our thesauri and ontologies.

Conclusion

The described technology of constructing ontologies for various domains was employed to create the so-called linguistic ontologies, which are used to solve problems of the automatic text processing. However, application of such technologies, connected with the processing of large text collections is also useful for the creation of ontologies in those domains, which are not directly connected with text processing.

The carried out analysis of the electronic text collection ensures:

completeness of concepts covering in reference to the collected corpus;

objectivity of concepts and terms interpretation, as various texts from the collection are analyzed.

"Minimal" relations set

makes possible to begin the ontology construction at once, as soon as a task is set and a domain is determined;

provides a conceptual basis for communicating with experts in the given domain;

provides the initial domain structuring which may be used as a basis for singling out special relations in the domain.

Acknowledgments

The work has been performed thanks to the support of the Russian Fund of Basic Research, grant № 02-07-90279.

Bibliography

[Loukachevitch, 2002] Loukachevitch Natalia V., Dobrov Boris V. Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool // Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002) / M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.) – Vol.1 – 2002, Gran Canaria, Spain – p.115-121.

[Loukachevitch N., 2002] Loukachevitch N.V., Dobrov B.V. Thesaurus of the Russian language for automatic processing of large text collections// Computer linguistics and intellectual technologies: Works of the International seminar Dialogue'2002 / edited by. a.S.Narinyani – м.: Nauka – 2002. – Vol.2 - p.338-346. In Russian.

[Nevzorova, 2001] Nevzorova O.A., Fedunov B E. System of analysis of technical texts "Lota": main conceptions and project decisions. // RAS publishing house. Theory and management systems – 2001. – № 3.– pp. 138-149. In Russian.

[Dobrov, 2002] Dobrov B.V., Loukachevitch N.V., Nevzorova O.A. Computer-aided construction of the applied ontology: technological aspects // International conference IEEE Artificial intelligence systems (IEEE AIS'02) Gelendzhik-Divnomorskoe, 5-10 September 2002 года – Text processing and cognitive technologies: Collection (ed. 7) / edited by V.D.Solovyev – Kazan: Otechestvo 2002. – pp.103-109. In Russian.

[Loukachevitch, 1996] Loukachevitch N.V. Computer-aided formation of information searching thesaurus of social and political life in Russia// NTI. ser.2. - 1995. - N 3. - pp.21-24. In Russian.

[Smith, 1998]Smith B. Basic tools of formal ontology. In "Formal Ontology in Information Systems", N. Guarino, ed.

[Guarino, 1998] Guarino N. Some Ontological Principles for Designing Upper Level Lexical Resources. In "Proceedings of First International Conference on Language Resources and Evaluation".

Author information

Boris Dobrov – Research Computing Centre Moscow State University, Russia, Moscow, Vorobjovy Gory; e-mail: dobroff@mail.cir.ru

Natalia Loukachevitch – Research Computing Centre Moscow State University, Russia, Moscow, Vorobjovy Gory; e-mail: louk@mail.cir.ru

Olga Nevzorova – Chebotarev Institute of Mathematics and Mechanics, Russia, Kazan, ul. Kremlevskaja, 18; e-mail: Olga.Nevzorova@ksu.ru.