A Conceptual Ontology for Machine Translation from/into Polish

Abstract

The paper presents a conceptual ontology that has been developed for the purpose of machine translation from and into Polish. The ontology has been applied in *Translatica* – a general domain MT system that currently translates text between Polish and English and aims at the development of other language pairs including Polish. *Translatica* ontology is designed mainly for disambiguation purposes and contains noun terms. The ontological concepts are applied as semantic values in lexical rules for verbs, adjectives and prepositions. The ontology is based on WordNet. The paper compares the assumed approach to those taken in other transfer-based and interlingua-based systems. It also points out and justifies the differences between the *Translatica* ontology and WordNet.

1. Introduction

Researchers distinguish three types of MT systems (Hutchins, Sommers, 1992):

- Systems that use terminological material but do not contain declarative knowledge bases of the domains they operate in
- Systems that use knowledge about concepts or facts for specific tasks like syntactic disambiguation or word sense disambiguation
- Systems that construct a deep meaning representation.

Translatica is a transfer MT system between Polish and English that falls upon the second group of the above classification. As mentioned in (Hahn, 2002) "systems need non-linguistic knowledge to solve a number of linguistics tasks". Although *Translatica* has no ambition to build an iterlingual representation of the meaning, it still needs non-linguistic knowledge for disambiguation purposes. Hahn distinguishes three types of non-linguistic knowledge used in MT systems: conceptual knowledge, world knowledge and facts, and situation knowledge. Aiming at general-purpose translation *Translatica* does not use world knowledge or situation knowledge. The conceptual knowledge is treated in the lexicon.

The lexicon of *Translatica* is based on large traditional dictionaries (PWN-Oxford, 2002), (PWN-Oxford, 2004). Human-readable dictionaries do not contain explicit conceptual knowledge – such knowledge is often implicitly delivered in examples of usage. Part of conceptual knowledge has been imported from the traditional dictionaries to the *Translatica* lexicon automatically (Jassem, 2003). Still, there is need for human processing to extract and formalize the knowledge that is left in the above-mentioned traditional dictionaries for human intuition and lingustic competence. A group of lexicographers have been working on the task since February 2003. The lexicographers need an ontology that would spread over all words of the general purpose as well as a set of well-defined hints on how to fit word senses to the concepts of the ontology.

The paper presents the general domain ontology, based on WordNet (Miller, 1995), used in *Translatica*. The paper lists the most general concepts of the ontology, underlines similarities and differences between this ontology and those used in other MT systems and presents examples how and why the conceptual hierarchy differs from that devised in WordNet.

2. Concept ontologies in MT systems

Concept ontologies are usually applied in machine translation that use "knowledge" that is in Knowledge-Based Machine Translation (KBMT) systems. The first systematic attempt was the KBMT-89 project (Mitamura, Nyberg, 1992) that aimed at delivering bi-directional translations of PC manuals for English and Japanese. The assumption behind the project was the use of Interlingua – a meaning representation that can serve for translations to a number of languages. The basic components of the system were: the ontology of concepts, lexicons and grammars for each language, and mapping rules between language-specific resources and Interlingua. The KBMT-89 ontology contained 5 basic concepts: object, event, property, relation, and attribute. Those basic concepts formed the roots for the semantic hierarchy.

The direction of further development of KBMT systems was the creation of a language-independent ontology that would serve to build a language-neutral interlingual format. In the Mikrokosmos project (Mahesh, 1996) the ontology acquired over 2000 concepts and reached the depth of 10 levels or more. The top nodes of the hierarchy were: object (with the subnodes: physical object, mental object, social object), event (with the subnodes: physical event, mental event, social event) and property (subnodes: attribute, relation). Although the authors allowed the hierarchy to acquire new concepts "the top levels of the hierarchy have proved very stable".

In recent years the idea of the interlingual ontology has been undertaken by the W3C consortium (Hahn, 2003). The ontology should consist of interlingual concepts (with English designators) that are not linked with specific words, and logical relations between them (like transitivity).

On the other pole are transfer-based systems. Their basic aim is to use conceptual ontologism mainly for disambiguation of word senses. The ALT-J/E system (Yamazaki, T., Pazzani, M., 1996) is a classical example. The semantic hierarchy includes only nouns. The highest levels of the hierarchy are Concrete (Agents, Places, Objects) and Abstract (Abstract Things, Things, Abstract Relationships). The ontological concepts are used in lexical translation rules that are either created by hand or learned from examples. The process of building and creating the ontology is parallel to the process of creating new translation rules – therefore it is crucial that the ontology should allow for updating. The resulting ontology of the system is 12 levels deep and has 790 intermediate nodes and 1925 leaf nodes.

The semantic hierarchy of the ALT-J/E system was built from scratch. It is more tempting to build an ontology

basing on an existing resource. In (Rigaud, Agirre, 1995) the authors examined the possibility of creating new ontologies by linking WordNet with bilingual dictionaries. Their experiments showed that this task couldn't be executed fully automatically. The reason is that the word-senses in WordNet and word-senses in bilingual dictionaries coincide only partially.

3. Creating *Translatica* ontology

The above review of MT ontologies shows that Interlingua systems (e.g. KBMT-89, Mikrokosmos) attempt to formulate ontologies that include noun, verb and adjective concepts. On the other hand, transfer-based MT systems (e.g. ALT J/E) need noun ontologies only. The concepts from noun ontologies are applied in translation rules for verbs, adjectives (prepositions) for disambiguation purposes.

Translatica follows the transfer-based approach. The ontology is created for nouns. Translation rules for other parts of speech apply the ontology and are included in the lexicon.

The *Translatica* lexicon follows the source/target approach. That means, "a word has as many senses as it has translation equivalents" (Baldwin, Hutchinson, Bond, 1999). A word and its translation equivalent will be further referred to as a word-sense.

Translatica ontology of noun concepts was created in the following steps:

1) Choosing an existing ontology as the initial resource of conceptual knowledge. WordNet (WN) seemed to be the best choice because of its availability as well as a sufficient density of concepts (in contrast to e.g. the Penman Upper Model (Bateman, 1990) that seemed to be too general for the task).

2) Using WN to assign semantic categories to nouns and noun phrases. Each noun (and noun phrase) was classified to a certain concept: one of its hypernyms according to WN or the noun (the head of the phrase) itself, on the basis of lexicographers' intuition.

3) Using WordNet concepts as semantic values for verb agents and themes. The concepts were chosen on the basis of examples of usage given in traditional dictionaries. For example, if two usages of the same verb-sense *break* had the themes: *window* and *glass* respectively, the lexicographer was supposed to find out the most accurate concept that subsumes both *window* and *glass*. A similar method was chosen for describing adjectives – lexicographers' task was to determine a concept that is modified by an adjective-sense. (Lexicographers' task was facilitated by a program that analyzed agents and themes in the examples against WordNet and provided lists of prompts to choose from).

4) Creating the first ontology on the basis of WordNet. The concepts were selected on simple statistics: those most frequently chosen by lexicographers were extracted from the WN hierarchy and formed the initial *Translatica* ontology. As a result, the ontology consisted of general concepts that were easy to understand and recognize by lexicographers. At that time *Translatica* ontology followed WN in the definition of the concepts. However, some intermediate nodes had been removed from the WN hierarchy. For example, in WN there exist intermediate nodes between ILLNESS and CONDITION (ILL HEALTH) and between INJURY and CONDITION (PATHOLOGICAL STATE). In *Translatica* both ILLNESS and INJURY are direct children nodes of CONDITION. We have found out that any sorts of ILL HEALTH or PATHOLOGICAL STATE may be classified either to ILLNESS (INJURY respectively) or CONDITION without any damage for the disambiguation process. The resulting ontology was 9 levels deep. The format of the ontology allowed for adding new nodes and deleting old ones if necessary.

5) Testing translation against the ontology

6) Analysing examples of poor semantic disambiguation. This is discussed in Section 4.

7) Adjusting the WN-based ontology in order to make it fit for machine translation from and into Polish. This is presented in Section 5.

4. Analysis of semantic disambiguation

In order to examine the process of semantic disambiguation on the basis of the designed ontology we analyzed translation of verb phrases and observed how the semantic categories assigned to agents and themes of verbs influence the quality of translation. We found out that poor translations resulted mainly from imprecise mappings of nouns to the concepts in the lexicon. We determined the following reasons of that undesirable state:

- the lack of some useful categories in the devised hierarchy (e.g. we found the need for new categories, eg. CLASH, VIEW)
- mapping of nouns to too broad concepts (a number of nouns were classified as belonging to a general category PSYCHOLOGICAL FEATURE instead of more specific categories)
- Organization of the hierarchy that did not help disambiguation of verb agents (e.g. SHOW was subsumed by COMMUNICATION although the sorts of SHOW cannot be said, read or written like other sorts of COMMUNICATION, but they can be organized and attended to like EVENT sorts).
- the imprecise definition of the word-sense as the word and its equivalent in the target language. This resulted in assigning several (and very often distant) semantic categories to word-senses (e.g. the word-sense *head-głowa* was assigned to both BODY PART and PERSON which resulted in an exemplary translation: *My head is aching* into *Moja głowa cierpi* (*My head is suffering*) because the sense of *head* was erroneously chosen as PERSON).

Due to the above problems we decided to re-define and re-organize the existing ontological hierarchy in such a way so that it provides an extensive and efficient semantic classification of nouns, and to re-divide word-senses in the lexicon so that each sense could be mapped to as small number of semantic categories as possible (desirably one).

5. New concepts and definitions in *Translatica* ontology

As mentioned above, the *Translatica* ontology originated from WordNet. This section lists and justifies adjustment of the WN noun ontology made for the purpose of machine translation from/into Polish.

5.1. Introduction of new categories

1) SOCIAL PHENOMENON category was added as a child node of PHENOMENON. For disambiguation purposes it seems important to distinguish certain sorts of PHENOMENA that affect SOCIAL GROUPs (e.g. society) or ORGANIZATIONs (e.g. market) from other PHENOMENA, either general (inflow, anomaly), or NATURAL (frost, eruption, flame). Unlike any other PHENOMENON, SOCIAL PHENOMENON can exceed a measure or a number, increase, rise or amount to a measure or a number, or drive another SOCIAL PHENOMENON (e.g. inflation). The SOCIAL PHENOMENON category also helps disambiguate meanings of some Polish verbs: przekroczyć is mapped to English exceed if the agent is a sort of a SOCIAL PHENOMENON, and to cross if the agent is PERSON or ANIMAL; the verb ograniczyć should be translated into curb if its theme is a sort of SOCIAL PHENOMENON, otherwise it should be rendered as limit/restrict (or surround, or cut down on).
2) VIEW has been added as a child node of PSYCHOLOGICAL FEATURE and defined as a *way of regarding things, or a personal belief or judgment.* VIEW differs from other PSYCHOLOGICAL FEATURE sorts in that can be propagated (in Polish: krzewione), professed, followed or believed in.
3) CLASH is a new child node of EVENT. One can provoke or fall in a CLASH but not another sort of EVENT. CLASH category also helps to distinguish meanings of some Polish verbs: wywołać translates to provoke if the theme is CLASH, otherwise it is mapped to the English equivalent trigger off or evoke.

5.2. Modifications of the hierarchy organization

1) In the WN hierarchy the top-level concepts (and children nodes of ANY) are: POSSESSION, EVENT, PHENOMENON, STATE, PSYCHOLOGICAL FEATURE, ACT, GROUP, ABSTRACTION and ENTITY. In the organization of the top-level concepts we decided to follow the ALT-J/E system, where ANY divides into CONCRETE (*Translatica:* OBJECT) and ABSTRACT (*Translatica: NONOBJECT*), OBJECT contains *3-dimensional* ontological sorts which one can see, touch, or feel. The children nodes of NONOBJECT are: POSSESSION, EVENT, PHENOMENON, STATE, PSYCHOLOGICAL FEATURE, ACT and ABSTRACTION. NONOBJECTs are known by intuition and reasoning, whereas OBJECTS are known be senses. NON-OBJECTS can be experienced, whereas OBJECTS cannot.
2) In the *Translatica* ontology SHOW (e.g. film, opera, performance, and concert) is-an EVENT rather than COMMUNICATION because one can attend a SHOW, go to it or see it. These verbs are characteristic for having EVENTs as their themes. One cannot attend to, go to or see a MESSAGE (which is a subconcept of COMMUNICATION). Moreover, there is a group of verbs for which SHOW is a typical agent (SHOW may amuse, engross, take off) or theme (one can cast, rehearse SHOW).
3) MONETARY UNIT, a child node of UNIT OF MEASUREMENT in WN is-a POSSESSION in *Translatica* because of the verbs that take both MONETARY UNIT and POSSESSION as their themes, such as pay, earn, save, invest.

5.3. Removal of categories not useful in MT disambiguation

QUANTITY has been removed from the Translatica hierarchy because it has a very similar sense to AMOUNT: *How there is of something that you can quantify.*
MUSICAL COMPOSITION has been deleted because very few nouns in the dictionary have been assigned this sense. The category proved not useful for word sense disambiguation since there is a very similar concept in the hierarchy: MUSIC.

5.4. Re-definition of ontological concepts

The primary aim of the ontology – application in machine translation from/into Polish – caused the need to re-define certain concepts.
The WN definition of ACTION is *something done (usually as opposed to something said),* ACTIVITY is just *any specific activity.* We suggest to regard ACTION as *something that a person does or causes to happen at a given place and time that is not repeated regularly* (e.g. abortion, depilation, voting, ethnic cleansing), and ACTIVITY as *something that a person does or causes to happen and that extends in time or is repeated regularly* (e.g. playing, acting, skiing, working, entertainment, censorship).
Re-definition sometimes results in re-organization of hierarchy of concepts. For example, in the *Translatica* ontology CRIME and GESTURE are children nodes of ACTION, whereas in WN they are subsumed by ACTIVITY.
SOCIAL RELATION has been defined as a RELATION between PERSONs or SOCIAL GROUPs. (In WN SOCIAL RELATION is a RELATION between LIVING THINGs). Our approach is consistent with that of Mikrokosmos, where the ontology provides a SOCIAL-OBJECT RELATION concept. Characteristic verbs that take SOCIAL RELATION as their theme are: enter into, or form (e.g. fraternity). SOCIAL RELATION can be established or broken. This category also helps distinguish meanings of some Polish verbs: zawrzeć translates into make if the object is a sort of SOCIAL RELATION, otherwise it is mapped to the English equivalent contain or conclude.
Table 1. visualizes some of the differences between the semantic hierarchies in WordNet and Translatica: the left column contains an excerpt of the WN ontology, the right column presents the tree of corresponding *Translatica* concepts.

6. Conclusions

Translatica ontology consists of 130 non-leaf concepts. The leafs of the ontology are noun-senses (words or phrases) extracted form traditional dictionaries (PWN-Oxford 2002), (PWN-Oxford, 2004). The ontology is designed for disambiguation purposes (rather than knowledge representation) in MT systems from/into Polish. The density of the ontology is a compromise between the need for accurateness of translation (that calls for great density) and domain generality (in order to develop lexical rules for the general domain lexicon in a

definite time period, the number of non-leaf concepts cannot be too high).
The ontology will be also applied in future (2005/2006) for the development of translation systems between Polish and Russian and German.

ANY	ANY
POSSESSION	NONOBJECT
EVENT	POSSESSION
SOUND	MONET. UNIT
PHENOMENON	EVENT
NATURAL	SHOW
PHENOMENON	CLASH
ACT	PHENOMENON
ACTION	NAT. PHEN.
ACTIVITY	SOUND
GAME	SOC. PHEN.
SPORT	ACT
DANCING	ACTION
POST	CRIME
WORK	GESTURE
CRIME	ACTIVITY
	GAME
	SPORT
	DANCING
	POST
	WORK

Table 1. Fragments of WordNet (left) and *Translatica* (right) ontologies

Bibliography

Baldwin, T., Hutchinson, B. Bond, F. (1999). A valency Dictionary Architecture for Machine Translation. In: *Eight International Conference on Theoretical and Methodological Issues in Machine Translation: TNI 99 (pp. 207-214)*, Chester

Bateman, A.J. (1990). Upper Modeling: A general organization of knowledge for natural language processing In *Proceedings of the 5th International Language Generation Workshop*, Pittsburgh

Hahn, W. (2003). Knowledge Representation in Machine Translation. In *Proceedings of EU Conference „Knowledge in Text and Translation (pp. 37 – 51)*, Aarchus

Hutchins, H. & Sommers J. (1992). *Introduction to machine translation.* London , Academic Press.

Jassem K. (2004). Applying Oxford-PWN English-Polish Dictionary to Machine Translation. In: *Proceedings of the Ninth EAMT workshop, Valetta, Malta (pp.98-105)*

Mahesh, K. (1996). Ontology Development for Machine Translation: Ideology and Methodology. Computing Research Labolatory MCCS–96–292. New Mexico State University,

Miller, G.A. (1995). WordNet: A lexical database for English. In *Communications of the ACM, 38(11), pp. 39 - 41* (!!!- chyba się strony nie zgadzają)

Mitamura, T., Nyberg, E.H. (1992). Hierarchical Lexical Structure And Interpretive Mapping In Machine Translation. In *Proceedings of COLING-92, Nantes, France*

Rigaud, G. Agirre E. (1995). Disambiguating bilingual nominal entries against WordNet, *Workshop On The Computational Lexicon - ESSLLI 95*

Yamazaki, T., Pazzani, M. (1994). A Cluster Analysis Approach to Learning a Semantic Hierarchy for Machine Translation. In *Proceedings of ML-COLT'94 Workshop on Constructive Induction and Change of Representation, (pp. 79-85)*

Yamazaki, T., Pazzani, M. (1996). Acquiring and Updating Hierarchical Knowledge for Machine Translation Based on a Clustering Technique. In *S. Wermter, E. Riloff, and G. Scheler, editors, Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, (pp. 329 – 342).* Springer-Verlag, Berlin, Germany.

PWN-Oxford (2002) Wielki słownik angielski-polsko, ed. Jadwiga Linde – Usiekniewicz. Wydawnictwo Naukowe PWN, Warszawa, 2002

PWN-Oxford (2004) Wielki słownik polsko-angielski, ed. Jadwiga Linde – Usiekniewicz. Wydawnictwo Naukowe PWN, Warszawa, 2004