ARCHIVES OF ACOUSTICS
32, 1, 75–87 (2007)

SEMANTIC DISAMBIGUATION IN AN MT SYSTEM BASED ON A BILINGUAL DICTIONARY

Krzysztof JASSEM[(1)], Agnieszka WAGNER[(2)]

[(1)] Adam Mickiewicz University
Department of Mathematics and Computer Science
Umultowska 87, 61-614 Poznań, Poland
e-mail: jassem@amu.edu.pl

[(2)] Adam Mickiewicz University
Institute of Linguistics
Międzychodzka 5, 60-371 Poznań, Poland
e-mail: wagner@amu.edu.pl

(*received October 16, 2006; accepted January 9, 2007*)

The paper presents an approach to semantic disambiguation in a transfer MT system based on a bilingual dictionary, i.e. the dictionary built from source-target pairs. The approach assumes building a concept ontology for nouns. The ontological concepts are applied as semantic values in lexical rules for verbs, adjectives and prepositions. A set of "semantic accommodation" rules is developed. The set is consulted during the process of semantic disambiguation that follows syntactical parsing. The approach has been applied in a commercial bi-directional Polish-English MT system called *Translatica*.

Keywords: machine translation, conceptual ontology, semantic hierarchy, semantic disambiguation, WordNet.

1. Introduction

As far as organization of the lexicon is concerned, transfer-based MT systems may be distinguished as either:

L1) Bilingual – in which entries are stored as source/target language pairs, or

L2) Monolingual – in which sense distinctions are specific for each language (BALDWIN, BOND and HUTCHINSON, [1]).

As regards the method of acquiring the lexicon, MT systems may be divided into:

A1) Hand-crafted – based on a "Hand-crafted Dictionary", built on a basis of traditional bilingual dictionaries;

A2) Learned – based on a "Learned Dictionary", where the transfer part of the dictionary is trained on bilingual text corpora (PINKHAM, SMETS, [13]).

With respect to the level of knowledge included in a translation process, the following types of MT systems can be distinguished:

K1) Terminological: Systems that use terminological material but do not contain declarative knowledge bases of the domains they operate in.

K2) Ontological: Systems that use knowledge about concepts or facts for specific tasks like syntactic disambiguation or word sense disambiguation.

K3) Deep: Systems that construct a deep meaning representation (HUTCHINS, SOMMERS, [4]).

HAHN [3] distinguishes three types of non-linguistic knowledge used in MT systems:

N1) Concept knowledge**;**

N2) World knowledge and facts;

N3) Situation knowledge.

Translatica is a transfer MT system between Polish and English with the aim of developing other language pairs including Polish. The system currently bases on existing traditional dictionaries (PWN-OUP 2002, PWN-OUP, 2004) and therefore assumes the monolingual hand-crafted approach (see also JASSEM, [5]). Aiming at general-purpose translation *Translatica* does not build a deep meaning representation and does not use knowledge about the world or current situation. Still, in order to solve lexical-semantic disambiguities the system needs some kind of concept knowledge. *Translatica* may thus be classified as:

- L1 (Bilingual),
- A1 (Hand-crafted),
- K2 (Ontological),
- N1 (Conceptual).

The paper presents an approach to semantic disambiguation that may be applied in an MT system of such a type.

2. A concept ontology

As mentioned above, the lexicon of *Translatica* is based on large traditional dictionaries [14, 15]. Human-readable dictionaries do not contain explicit concept knowledge – such knowledge is often implicitly delivered in examples of usage. A part of concept knowledge has been imported from the traditional dictionaries to the *Translatica* lexicon automatically (JASSEM, [5]). Still, there is a need for human processing to extract and formalize the knowledge that is left in the above-mentioned traditional dictionaries for human intuition and linguistic competence. A group of lexicographers have been working on the task since February 2003. The lexicographers need an ontology that would cover 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.

Section 2 presents a general domain ontology, based on WordNet (MILLER, [9]), used in *Translatica*. The paper lists some general concepts of the ontology and under-

lines similarities and differences between this ontology and those used in other MT systems. Moreover, some examples are given that show how and why the concept hierarchy differs from that devised in WordNet and SENSUS (KNIGHT, LUK, [7]).

2.1. 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, [11]) 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 in 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 top-level nodes of 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, [8]) 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" (MAHESH, [8]).

In recent years the idea of the interlingual ontology has been undertaken by the W3C consortium (HAHN, [3]). The ontology should consist of interlingual concepts (with English designators) that are not linked with specific words, and logical relations between them (like transitivity).

Transfer-based systems take a different approach. Their basic aim is to use concept knowledge mainly for the purpose of semantic disambiguation and selection of translation candidates. This can be achieved by applying an existing ontology, e.g. WordNet or SENSUS.

WordNet is a hierarchically organized lexical database that includes about 150 000 word senses: nouns, adjectives and verbs. The WN hierarchy is a tree-like structure (in fact, the WN branches intersect forming a structure of a lattice) where the top-level nodes of the tree represent concepts and low-level nodes – instances of concepts. Nouns are characterized by *superordinate terms* (*hypernyms*) and sets of distinguishing features. Relations between nouns include: *synonymy* (e.g. TOOL = INSTRUMENT), *hypernymy* (ARTIFACT is a hypernym for INSTRUMENTALITY which in turn is a hypernym for IMPLEMENT, the hypernym for TOOL) and *hyponymy* (any of {comb, drill, plow, power tool, rake} is-a TOOL). A *superordinate term* (*hypernym*) for a concept constitutes its parent node in the hierarchy and a *hyponym* forms a daughter node. *Synonyms* of a concept are grouped into synsets. In this way nouns are organized in

a *lexical inheritance system*: *A system in which each word inherits the distinguishing features of all its superordinates* (MILLER, [10]). The concept knowledge is organized in WN in such a way as to prove useful for semantic disambiguation in NLP systems, including MT systems.

RIGAUD, AGIRRE, [16] examined the possibility of creating ontologies for new languages – by linking WordNet with bilingual dictionaries. Their experiments showed that this task couldn't be executed fully automatically. The reason is that word-senses in WordNet and word-senses in bilingual dictionaries coincide only partially.

The SENSUS ontology (KNIGHT, LUK, [7]) has been created as a result of merging several sources of concept knowledge:

- The PENMAN Upper Model (see BATEMAN, [2]) and ONTOS (see NIRENBURG, DEFRISE, [12]) – high–level ontologies, which provided top–level nodes in the hierarchy (general concepts),
- a list of Semantic Categories (extracted from variuos dictionaries),
- WordNet,
- Dictionaries: monolingual English, English–Japanese and English–Spanish dictionaries, which provided intermediate and low–level nodes (specific concepts).

The resulting ontology consists of about 70 000 concepts and can serve as a basis for construction of a domain-specific ontology. In the organization of concept knowledge SENSUS follows WordNet: Noun concepts are related by *synonymy*, *hypernymy* and *hyponymy*.

While WordNet is not directed to be used in a specific domain of NLP, the SENSUS ontology has been devised primarily for use in MT systems. In order to be applicable in domain-specific machine translation the SENSUS ontology allows for extension: The SENSUS paradigm says that new concepts, both general and specific, should be easily added to the hierarchy or merged with the existing concepts. On the basis of the general-domain ontology it should be possible to develop domain-specific ontologies (e.g. for military air campaign planning). However, while expanding the ontology, the principles of the organization of the structure, defined by the developers, should be followed in order to keep the ontology coherent. SENSUS is the source of concept knowledge in the GAZELLE machine translation system (developed in the Information Sciences Institute, University of South California) that translates from Japanese, Arabic and Spanish into English.

Some MT systems design their own ontologies from scratch. The ALT-J/E system (YAMAZAKI, PAZZANI, [18]) is an example. This is a transfer-based system that uses semantic hierarchy of noun concepts. 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.

2.2. Translatica *ontology*

It may be concluded that interlingua systems (e.g. KBMT-89, Mikrokosmos) tend to formulate ontologies that include noun, verb and adjective concepts, whereas transfer-based MT systems (e.g. ALT J/E) need only noun ontologies and apply noun concepts in translation rules for verbs, adjectives (and 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.

Translatica ontology of noun concepts was created in the following steps (more details are given in JASSEM, WAGNER, [6]):

1) Choosing WordNet as the initial resource of concept knowledge because of its availability as well as a sufficient density of concepts.

2) Using WN to assign WordNet semantic categories to nouns and noun phrases (This task was executed semi-automatically: a lexicographer was expected to chose from a set of prompted categories – WN hypernyms of the noun (the head of the phrase) or the noun itself).

3) Using WordNet concepts as semantic values in lexical-semantic rules created for verbs, adjectives and prepositions. The concepts were used to characterize agents and themes of verbs, nouns that are modified by adjectives or nouns that are linked by prepositions.

4) Creating the initial ontology on the basis of WordNet. The concepts were selected with respect to a simple statistics: Those most frequently occurring in lexical rules 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 who created rules. 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) Analyzing examples of poor semantic disambiguation on the basis of the assumed ontology. This is discussed in Sec. 2.3.

7) Adjusting the WN-based ontology for the purposes of machine translation from and into Polish. This is presented in Sec. 2.4.

2.3. Analysis of misdisambiguation

The concept ontology discussed in the previous section formed the basis for lexical rules consulted in the process of lexical-semantic disambiguation. The algorithm of disambiguation is discussed in Sec. 3.3. Here, we discuss the cases of wrong disambiguation that resulted from the initial ontology of concepts (Section 2.4 describes modifications to the ontology that improved the quality of disambiguation).

We found out that cases of poor translation resulted mainly from imprecise mappings of nouns to the concepts in the lexicon. We determined the following reasons of that undesirable state:

- lack of some useful categories in the devised hierarchy (such as SOCIAL_RELATION or VIEW),
- mapping of nouns to too broad concepts (a number of nouns were classified as belonging to a general concept PSYCHOLOGICAL FEATURE instead of more specific categories, e.g.: IDEA, VIEW, FEELING, ATTITUDE),
- organization of the hierarchy that did not always help disambiguation of verb agents and themes (e.g. SHOW was subsumed by COMMUNICATION although instances of SHOW cannot be said, read or written, but they can be organized and attended to like instances of EVENT),
- the imprecise definition of the word-sense as the word and its equivalent in the target language, which 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 a wrong 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 it has been decided to re-define and re-organize the existing ontological hierarchy in such a way as to provide 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).

2.4. New concepts and definitions in Translatica ontology

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

Deletion of WN concepts

Some concept-nodes in the WN hierarchy did not seem useful for disambiguation in translation from/into Polish – such nodes were deleted. The purposes of removing less useful concepts were:

- to facilitate the task of creating lexical rules (the fewer concepts there are, the easier task to find the appropriate concept for a rule),
- to ensure homogeneity of lexical rules (if some concepts differ only slightly, they may be mixed up in lexical rules).

• WN has intermediate nodes between ILLNESS and CONDITION (ILL HEALTH) and between INJURY and CONDITION (PATHOLOGICAL STATE). In *Translatica* both ILLNESS and INJURY are direct daughter nodes of CONDITION. It has been verified that any instance of ILL HEALTH or PATHOLOGICAL STATE may be classi-

fied either to ILLNESS (INJURY respectively) or CONDITION without any damage for the disambiguation process.

• QUANTITY has been removed from the *Translatica* hierarchy because it has a very similar sense to AMOUNT: *How much there is of something that you can quantify*.

• MUSICAL COMPOSITION has been deleted because the category proved not useful for word sense disambiguation since there is a very similar concept in the hierarchy, namely MUSIC.

• The CONSTRUCTION concept (a daughter node of NATURAL_LANGUAGE and COMMUNICATION in WN) has been removed in *Translatica*. The concept was of little use as far as disambiguation of lexical senses is considered: e.g. Polish verbs that take instances of CONSTRUCTION as their arguments can also take as arguments other instances of COMMUNICATION and still are mapped to the same English equivalents (e.g. the Polish verb zanotować translates into to write down whenever the theme is an instance of either CONSTRUCTION (e.g. expression, phrase, clause) or COMMUNICATION (e.g. word, information, message, address, note).

Introduction of new categories

• TIME_MOMENT (e.g. tomorrow, Monday, noon, date, turning point) is a new concept in the *Translatica* hierarchy, distinct from TIME_PERIOD (e.g. decade, spring, future, past, hour, and month), which is defined as *a particular point in time*. Instances of TIME_MOMENT are characteristic agents of verbs such as draw near (in Polish nadciągać) or get on for (in Polish dochodzić do), and themes of verbs like while away (in Polish: umilić), chat away (in Polish przegadać), be in effect (since), postpone (till). At the same time, these verbs do not take instances of TIME_PERIOD as their agents or themes. The preposition ***in*** collocates with TIME-PERIOD (and is translated into Polish as ***za***), e.g. 'in a month' and it does not link with TIME MOMENT (e.g. in Monday*).

• SOCIAL PHENOMENON has been added as a daughter node of PHENOMENON. VIEW has been added as a daughter node of PSYCHOLOGICAL FEATURE and defined as *a way of regarding things*, or *a personal belief* or *judgment*. CLASH is a new child node of EVENT.

Modifications of the hierarchy organization

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. However, in the organization of the top-level concepts *Translatica* follows the ALT-J/E system, where ANY divides into CONCRETE (*Translatica*: OBJECT) and ABSTRACT (*Translatica*: NONOBJECT). The OBJECT class contains 3-dimensional ontological sorts that one can see, touch, or feel. The daughter 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, but OBJECTS cannot.

In the *Translatica* ontology SHOW (e.g. film, opera, performance, concert) is-an EVENT, whereas in the WN it is subsumed by COMMUNICATION. MONETARY UNIT, a daughter node of UNIT OF MEASUREMENT in WN, is-a POSSESSION in *Translatica*. KNOWLEDGE DOMAIN (e.g. ethics, philosophy, physics, anatomy) is a subconcept of PSYCHOLOGICAL FEATURE in WN, but in *Translatica* it is-an ABSTRACTION (for discussion on those changes (see JASSEM, WAGNER, [6]).

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 WN, CRIME (e.g. theft, murder) is subsumed by ACTIVITY, whereas in the Translatica ontology it is-an ACTION, because it occurs only once at a given time and place.

• 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 themes are enter into or form (e.g. fraternity), abolish (e.g. slavery, dictatorship). SOCIAL RELATION can be established or broken. Something may cause SOCIAL_RELATION (e.g. relationship) to cool off. The SOCIAL_RELATION category helps disambiguate meanings of some Polish verbs. For example the verb zawrzeć translates into to make if the object is-a SOCIAL RELATION, otherwise it is mapped to the English equivalent to contain or to conclude. The verb objąć has the sense of to take (up) if the theme is an instance of SOCIAL_RELATION (e.g. patronage, protectorate, power, command, leadership); otherwise it means to embrace (e.g. PERSON), to assume (e.g. power, POST), or to grasp (e.g. a sort of IDEA or VIEW).

• In the WN and SENSUS ontologies INSTRUMENT is defined as *DEVICE that requires skill for proper use and DEVICE is-an ARTIFACT invented for a particular purpose*. In *Translatica* INSTRUMENT (e.g. hammer, spade, knife, grind, pin) is *a non-power tool or a simple piece of equipment, that one can hold in hands in order to use for ACTIVITY or ACTION*. In *Translatica* DEVICE is defined in a similar way as MACHINE in the WN and SENSUS: It is-an *ARTIFACT powered by SUBSTANCE* (e.g. fuel) or *ARTIFACT that transmits or modifies energy to perform or assist in the performance of human tasks* (*ACTIVITIES* or *ACTIONS*).

The re-definition of INSTRUMENT and DEVICE resulted in that they are both direct daughter nodes of ARTIFACT and occur at the same level in the semantic hierarchy.

INSTRUMENT no longer is-a DEVICE. The following arguments justify the change: DEVICE can function, work and operate (in Polish działać) – this cannot be said of INSTRUMENT; and one can start DEVICE, but not INSTRUMENT. One can stick INSTRUMENT (e.g. pole, needle), cut, slash with it (e.g. knife, blade), dig with it (e.g. shovel, spade), sink it, sharpen it, but this does not apply to DEVICE. Unlike INSTRUMENT, DEVICE can be powered, charged, turned on/off.

Table 1 displays some of the differences between the semantic hierarchies in WordNet, SENSUS and *Translatica*: The left column contains an excerpt of the SENSUS ontology, the middle column shows the corresponding concepts in WordNet, and the right column presents the tree of corresponding *Translatica* concepts.

Table 1. Basic ontological concepts in SENSUS, WordNet and Translatica – excerpt.

SENSUS	WordNet	Translatica
OBJECT-THING	ANY	ANY
PROCESS		NONOBJECT
PHENOMENON	PHENOMENON	PHENOMENON
ECONOMIC PROCESS	ECON. PROCESS	SOCIAL PHEN.
NATURE PHEN.	NATURAL PHEN.	NATURAL PHEN.
		SOUND
EVENT	EVENT	EVENT
SOUND	SOUND	SHOW
OBJECT		CLASH
ABSTRACTION		
PSYCHOLOGIC. FEAT.	PSYCHOLOGIC. FEAT.	PSYCHOLOGIC. FEAT.
KNOWLEDGE	KNOWLEDGE DOM.	
IDEA	IDEA	IDEA
BELIEF	BELIEF	VIEW
ATTITUDE	ATTITUDE	ATTITUDE
FEELING	FEELING	FEELING
POSSESSION	POSSESSION	POSSESSION
QUALITY	ABSTRACTION	ABSTRACTION
INTERPERS. THING	ATTRIBUTE	ATTRIBUTE
TEXTUAL THING	QUALITY	KNOWLEDGE DOM.

3. Rule-based lexical-semantic disambiguation

In SANFILIPPO, STEINBERGER, [19], a method of using a monolingual thesaurus as a lexical database and a source of concept knowledge is proposed: Semantic disambiguation is based on linking senses of words from a bilingual dictionary (provided usually in the form of usage examples) to their senses stored in a monolingual thesaurus.

This is executed in three steps: First, each sense of the source language word from the bilingual dictionary is mapped to appropriate sense in the thesaurus and a set of all its synonyms is retrieved from the thesaurus.

In the next step, the synonyms are mapped to their equivalents in the target language (provided in the bilingual dictionary). The final choice of the best candidate is based on the frequency rates of equivalents.

The approach suggested here differs: Having at the disposal traditional dictionaries – adapted for MT purposes – it looks unjustified to calculate frequency of equivalents; why not to trust the research (intuition?) of the authors of traditional dictionaries who list the synonyms in a precise order? Instead, an MT system should concentrate on selecting the best sense of a word in a context.

The algorithm of semantic disambiguation suggested here consults two resources: a set of semantic-lexical rules and a set of semantic-syntactical rules.

3.1. Semantic-lexical rules

In *Translatica* semantic-lexical rules are stored in the lexicon. The rules take into account:

- semantic characteristics of agents and themes for verbs; these characteristics are expressed by means of the *Translatica* concept ontology,
- semantic characteristics of modified nouns for adjectives,
- semantic characteristics of nouns that form PPs together with a given preposition, for prepositions.

For example, the basic senses of the adjective *single* are described by Polish equivalents in the following way:

pojedynczy for any unspecified OBJECT (*a single house*),
jeden for any EVENT or ACT (*a single performance*),
stanu wolnego for PERSON (*a single woman*).

Disambiguation of noun senses is based (mainly) on examples of usage that are stored in *Translatica* lexicon of word phrases. For example, *to win a match* disambiguates a sense of *match* as an event rather than an artifact.

3.2. Semantic-syntactical rules

Semantic-syntactical rules define the way in which the semantic value of a syntactical component should be calculated. Semantic values are expressed in the form of subsets of all entities that form the system ontology. Leaves of the semantic hierarchy are single elements and higher-level nodes form non-trivial subsets. For example, a semantic value {'monkey', person} means all instances that are either human or 'monkey'.

The classical set operators are allowed for semantic values, i.e. negation (-{'monkey', person} defines all entities that are neither human nor 'monkey'), intersection (e.g. {'monkey', person} $\cap$ {worker} = {worker}) and sum (e.g. {'monkey', person} $\cup$ {worker} = {'monkey', person}).

Two approaches may be taken for semantic-syntactical rules:

1. The semantic-syntactical rules may be linked with syntactical rules, e.g.
   AdjPhr $\rightarrow$ Adjective NounPhrase
   Syntactical constraints
   AdjPhr := Adjective.Sem $\wedge$ NounPhrase.Sem.

2. The semantic-syntactical rules may be separated from syntactical rules, e.g.
{ AP //sem-syn rule for any AP component
AP := A.Sem ^ NP.Sem
}

The latter approach has two advantages:

- convenience: the same semantic-syntactical rule need not be re-written for all syntactical rules that result in a AP component,
- economy: during calculations syntactical analysis need not take into account all semantic senses of words.

More details on the format of the syntactical rules are intended for a camera-ready version of the paper.

3.3. The algorithm for semantic disambiguation

So far, three algorithms have been tested:

1. Semantic-syntactical rules determine the choice of all right-hand components of a rule.

Suppose that the following entries are included in the lexicon:

Source	**Sem Value of a Modifee**	**Target**
single	object	pojedynczy
	act, event	jeden
	person	stanu wolnego

Source	**Sem Value**	**Target**
match	instrument	zapałka
	event	mecz
	person	partia

Then, applying the rule:
$AP := A.Sem^{\wedge}NP.Sem$
to the sentence:
He did not win a single match
the obtained translation would be:
Nie wygrał partii stanu wolnego (back to English: *he did not win an unmarried person who is a good match*).

2. Semantic-syntactical rules determine the choice of all right-hand components but the senses of nouns are restricted by contextual or lexical-syntactical rules.

A restrictive contextual rule for the pair *match–partia* could be "marriage topic" (this is not the case in the *Translatica* lexicon); a restrictive lexical-syntactical rule (such rule is included in the *Translatica* lexicon) says that *match* is mapped into *partia* only if it is followed by a PP: for PERSON.

This method gives the following translation:
Nie wygrał pojedynczej zapałki (back to English: *He did not win an instrument used for lighting fire*).

3. Semantic-lexical rules determine the choice of selected right-hand components. Then, the above rule would have the form:

$AP := A^*.Sem^{\wedge}NP.Sem$

(the star denotes the element that undergoes disambiguation).

With such a solution the rule does not disambiguate the noun sense. The sense of the noun – as an EVENT – is disambiguated by means of an example of usage (to win a match) and the semantic rule assures the choice for the equivalent of the adjective: jeden. This gives a desired (though still not perfect) translation of the sentence: *Nie wygrał jednego meczu*.

Not surprisingly, Algorithm 3 has been eventually chosen for semantic disambiguation in *Translatica*.

4. Conclusion

The paper presents an algorithm for semantic disambiguation in an MT system that uses a bilingual, hand-crafted dictionary with the lexical-semantic rules based on a concept ontology.

The algorithm has been applied in *Translatica*, the commercial system that translates between Polish and three other languages: English, Russian and German.

References

[1] BALDWIN T., HUTCHINSON B., BOND F., *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 1999.

[2] BATEMAN A. J., *Upper Modeling: A general organization of knowledge for natural language processing*, Proceedings of the 5th International Language Generation Workshop, Pittsburgh 1990.

[3] HAHN W., *Knowledge Representation in Machine Translation*, Proceedings of EU Conference "Knowledge in Text and Translation", pp. 37–51, Aarchus 2003.

[4] HUTCHINS H., SOMMERS J., *Introduction to machine translation*, Academic Press, London 1992.

[5] JASSEM K., *Applying Oxford-PWN English-Polish dictionary to machine translation*, Proceedings of the Ninth EAMT workshop, pp. 98–05, Valetta, Malta 2004.

[6] JASSEM K., WAGNER A., *Conceptual ontology for machine translation from/into Polish*, The Language Technology Conference, Poznań 2005.

[7] KNIGHT K., LUK S., *Building a large-scale knowledge base for machine translation*, Proceedings of the American Association of Artificial Intelligence AAAI-94, Seattle 1994.

[8] MAHESH K., *Ontology development for machine translation: Ideology and methodology*, Computing Research Labolatory MCCS–96–292, New Mexico State University, 1996.

[9] MILLER G. A., *WordNet: A lexical database for English*, Communications of the ACM, **38**, 11, 39–41 (1995).

[10] MILLER G. A., *Nouns in WordNet: a lexical inheritance system*, International Journal of Lexicography, **3**, 4, 245–264 (1999).

[11] MITAMURA T., NYBERG E. H., *Hierarchical lexical structure and interpretive mapping in machine translation*, Proceedings of COLING-92, Nantes, France 1992.

[12] NIRENBURG S., DEFRISE C., *Application-oriented computational semantics*, [in:] *Computational Linguistics and Formal Semantics*, Johnson R. and Johnson M. [Eds.], Cambridge University Press, Cambridge 1992.

[13] PINKHAM J., SMETS M., *Modular MT with a learned bilingual dictionary: rapid deployment*, Proceedings of the International Conference on Computational Linguistics, 2002.

[14] PWN-OUP *Wielki słownik angielsko-polski*, Jadwiga Linde–Usiekniewicz [Ed.], Wydawnictwo Naukowe PWN, Warszawa 2002.

[15] PWN-OUP *Wielki słownik polsko-angielski*, Jadwiga Linde–Usiekniewicz [Ed.], Wydawnictwo Naukowe PWN, Warszawa 2004.

[16] RIGAUD G., AGIRRE E., *Disambiguating bilingual nominal entries against WordNet*, Workshop on the Computational Lexicon – ESSLLI 95, 1995.

[17] YAMAZAKI T., PAZZANI M., *A cluster analysis approach to learning a semantic hierarchy for machine translation*, Proceedings of ML-COLT'94 Workshop on Constructive Induction and Change of Representation, pp. 79-85, 1994.

[18] YAMAZAKI T., PAZZANI M., *Acquiring and updating hierarchical knowledge for machine translation based on a clustering technique*, [in:] *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, S. Wermter, E. Riloff, and G. Scheler [Eds.], pp. 329–342, Springer-Verlag, Berlin, Germany 1996.

[19] SANFILIPPO A., STEINBERGER R., *Automatic selection and ranking of translation candidates*, Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 200–207, Santa Fe, New Mexico, USA 1997.