An Approach to Building a Multilingual Translation Dictionary that Contains Case, Prepositional and Ontological Information

Maxim Mozgovoy
University of Aizu
Aizu-Wakamatsu, Japan
mozgovoy@u-aizu.ac.jp

Tuomo Kakkonen
University of Joensuu
Joensuu, Finland
tkakkone@cs.joensuu.fi

Abstract

In the multilingual reality of the modern world the task of translational electronic dictionaries development remains highly important. Many existing computer-based dictionaries lack the following features: (a) the availability of both machine-readable and human-readable formats; (b) strict formal definitions of phrase contexts that mandate specific translations. The current work addresses these issues by suggesting a new approach that incorporates formal definitions of a phrase context that is based on EuroWordNet-style ontological information. The proposed dictionary will enable the selection of the correct translation of a word in a given context, and will thus be suitable both for human use and NLP tasks.

1 Introduction

Because most dictionaries do not use a formal and consistent method for defining their words, their definitions are not convenient for using in *natural language processing* (NLP) algorithms. Even ordinary human users of a dictionary often encounter difficulties when they try to find a proper word card section (i.e. translation) for a polysemic word when translating from one language to another. This problem is aggravated when a word has a number of possible meanings and that are used in a variety of contexts. English prepositions are a class of words that clearly exhibit just such difficulties.

One way of solving this problem is to use a formal and invariable method of composing word cards that makes them both clearly organized and machine-readable. While such a procedure would be useful in several kinds of dictionaries, they are especially valuable for the kind of multilingual dictionaries that translators use for finding the correct translation of a particular word or phrase across several languages.

The purpose of this article is to propose a particular approach to the construction of machine-readable *multilingual translation dictionaries* (MTDs). This approach depends on being able to define words in monolingual dictionaries by using *concept types* and by linking them with a *translation table*. In this paper we will use English, Russian and Finnish to create examples that will illustrate this technique, because these particular languages enable us to demonstrate with clarity the concepts and procedures we wish to define. Since these languages also represent different groups of languages, they enable us to demonstrate how such an approach is generalizable to many types of languages.

2 Related Work

This work follows Prof. Tusov's approach [Tusov, 2004], initially proposed for parsing Russian texts. In its turn, Tusov's approach was inspired by Mel'čhuk's *Meaning-Text model* [Mel'čuk, 1995], and specifically by his concept of an *explanatory combinatorial dictionary* (ECD). But since Mel'čhuk's dictionaries are designed specifically for human use, their structure is not fully compatible with NLP applications and automatic text parsers. It is also worth noting that the construction of such a dictionary requires enormous human resources. Thus, for example, the four volumes of ECD for French took twenty years to construct although they contain only about 500 words [Mel'čuk et al., 1999].

Tusov's syntactic parser is based on a machine-readable dictionary, specifically designed to reflect a great deal of rich semantic information for each entry, both at the level of individual words, and at the level of real-world ontology. The definition of each word includes information about its possible *arguments* (a generalized idea of the well-known concept of *valency*, already incorporated in electronic dictionaries [Baldwin et al., 1999]). The inclusion of ontological information about individual words is also widely used in semantic-rich computer dictionaries such as EuroWordNet [Vossen et al., 1998]. Our proposed approach does not require a specific ontology. A number of existing ontologies can be combined for future use in a single conceptual map [Knight and Luk, 1994].

3 Design Principles

Most of the problems connected to the use of dictionaries arise out of the fact that words in different languages have multiple, partly intersecting meanings. The English word *"play"*, for example, can be used to refer to at least two completely different concepts, as in *to play a game* and *to play a musical instrument*. It is simple to translate an English phrase with the word *"play"* into Russian because Russian also uses the word *play* ([igrat']) in both these contexts. In Finnish, however, there are three different words to express what in English is expressed by the single word *"play"*. Thus, one uses the Finnish word *"pelata"* to express the idea of playing a game, the word *"soittaa"* to express the idea of playing an instrument or of making a telephone call, and the word *"leikkiä"* to express the idea of playing childishly (i.e. not seriously). It is up to the translator to select the correct word by taking note of the semantic context. A dictionary therefore needs to explain the ontological context expressed by both the source and the target language. Context, in turn, is defined by a sequence of words that depend upon one another and the word for which a meaning is being determined.

In our model, the three possible contexts for the Finnish equivalent of "to play" can be described in the following formal way:

PLAY_1(*game*|*object-not-instrument*)
    pelata
PLAY_2(*instrument*|*phone*)
    soittaa
PLAY_3(*empty*|*child-game*)
    leikkiä

In this case, *game*, *instrument*, *child-game*, and *object-not-instrument* are not actual words: they are *concept classes* that denote the "type" of a word that can be used in the corresponding alternative. In this model, we treat words such as *"piano"* and *"violin"* as representatives of the concept class *instrument*, while *"chess"* and *"checkers"* are representatives of the class *game*.

3.1 Hierarchy of Concepts

A Finn, for example, would distinguish several versions of the word *"play"* because the word is applicable to several game names in the "abstract" class *game*, and also so to the names of single instrument names in a class *instrument*. A possible (though challenging) solution to the problem of defining a concept hierarchy for translation dictionaries would be to construct a universal tree of concepts that would include classes such as *game*, *instrument*, and all the other objects of human life. Arguably, such a classification could be developed for a single language since all the users of that language use it implicitly in their language constructions. It should be noted that the purpose of such a hierarchy is to reflect the linguistic reality of a particular language's world (so it could be used for the purpose of syntactic parsing), and not to provide a universal taxonomy of objects in philosophical sense.

But the problem of constructing a hierarchy of concepts becomes far more challenging when one constructs a multilingual dictionary. The complexity of this challenge is caused by the fact that there are words or even whole concept classes that exist in one language realm, but which are absent in another[1]. The construction of a joint concepts hierarchy is not a simple task even when one is using only two languages. Fortunately, such ontologies (such as those of [Vossen et al., 1998] and [BMIR, 2008]) do already exist, and it is possible to reuse them.

We decided to base our concept classes on the *Top Ontology of EuroWordNet* [Vossen et al., 1998], because it offers a concise and tested hierarchy of concepts which has been used as the basis for describing dictionaries in several European languages. The ontology is divided into (1) concrete entities (1st order), (2) entities that are not physical things but which can be located in time and that can occur or take place sporadically rather than exist continuously (2nd order), and (3) physically unobservable propositions that exist independently of time and space but which can be inferred.

Since individual concept classes can exist independently of a single unified structure, one might question the need for a *hierarchical organization* of the concepts. The hierarchical

[1] Although the discussion of this question is beyond the scope of this work, it is worth noting that in any particular case a translator has to locate a source (to be translated) word somewhere in the world of the destination language. In the simplest case, foreign words are left "as is", like Russian *"samovar"* and Finnish *"sauna"*.

structure is necessary because certain words sometimes need to be regarded as representatives of their own subclasses, while on other occasions they need to be treated as members of a single superclass. One only needs to revisit the example above to see the truth of this. In that example we saw that Finnish uses three different words to represent the various meanings of the English word “play”, depending on whether one is playing a game, or whether the playing is “serious”, or whether is an instrument that is being played. However, a Finnish verb “pitää” (to like) treats all objects equally: we can thus like both a “viulu” (violin) and a “jalkapallo” (football). Thus, there is a need for a joint type that contains both a “viulu” and a “jalkapallo”.

3.2 Linking the words

The kind of dictionary we have been describing above helps a translator to select *which* word to use in a translation, but it offers no help in showing the translator *how* to use it. Both the word in the source language and the word in the target language might, for example, require specific kinds of links with its dependent words (such as its subject and the object), and these are usually established by means of a specific preposition or a grammatical case.

Consider, for example, the following definitions for the English word *“to travel”*:

ST/Dynamic TRAVEL_1(*vehicle*)
to travel using a specified vehicle
ST/Dynamic TRAVEL_2(*place*)
to travel to a specified place

Here ST/Dynamic is a concept class name of the word *travel* itself.

What these definitions do not reveal is that, in the first case, it is necessary to use the preposition *“by”*, while, in the second case, the preposition *“to”* is required. If one were to translate such a construction into a target language, one might very well encounter corresponding differences in word link types there. It is therefore necessary to be able to describe the links explicitly.

The languages that we have been using for our examples reveal clear differences in their methods of constructing dependencies between words. Modern English uses mostly preposition-based links: the type of action represented in the phrases *“to travel by car“* and *“to travel to Moscow”* is defined and indicated by the use of a specific preposition. It is therefore unsurprising to observe that English utilizes an extensive repertoire of prepositions, and that most prepositions have multiple meanings.

Finnish, by contrast, represents another extreme: it is based on an agglutinative, inflectional case-driven model. In the example using *“travel”* in the paragraph above, it would be necessary in Finnish to modify the forms of the words *“car”* and *“Moscow”* without introducing any prepositions in order to indicate the proposed actions. Finnish consequently makes use of an extensive case system, and all Finnish nouns have 15 cases. Russian represents a language that make use of combinations of prepositions and inflections. Each of these languages possesses a moderate number of prepositions (compared, for example, to English), and the cases of nouns are indicated by appropriate inflections of the nouns concerned. These languages therefore require the use of both the correct preposition and its corresponding word form (inflection).

Establishing a prepositional word link in the dictionary is a straightforward matter. Instead of an immediate word class of an object, we specify the exact preposition that must be used:

ST/Dynamic TRAVEL_1(BY_1)
to travel using a specified vehicle
ST/Dynamic TRAVEL_2(BY_2)
to travel across something
ST/Dynamic TRAVEL_3(TO_1)
to travel to a specified place
Preposition BY_1(*vehicle*)
a vehicle used to travel (e.g. car)
Preposition BY_2(*place*)
something we go across
(e.g. “travel by riverside”)
Preposition TO_1(*place*)
place we travel to (e.g. “Moscow”)

Describing a case link requires additional notation, as is shown in the following Finnish example:

ST/Dynamic
TRAVEL_1(*vehicle*: adessive_case)
(“matkustaa autolla”)
(to travel by car)
ST/Dynamic
TRAVEL_2(*place*: partitive_case)
(“matkustaa joen vartta”)
(to travel by riverside)
ST/Dynamic
TRAVEL_3(*place*: illative_case)
(“matkustaa Helsinkiin”)
(to travel to Helsinki)

ST/Dynamic
TRAVEL_4(*place*: allative_case)
("matkustaa Tampereelle")
(to travel to Tampere[2])

3.3 Idioms and Compound Words

Some languages tend to create new words by combining the stems of multiple words (i.e. they create compound words), while other languages keep the initial words separate and form a word combination. Finnish and English are good representatives of these families. For example, the English "bald eagle" is represented by just one word in Finnish: "valkopäämerikotka" (lit. "white-headed sea eagle"). When translating a text from English into Finnish (or another language with a similar characteristic), it becomes necessary to be able to identify phrases in the source language that can be translated with a single word in the target language, and vice versa.

We can express this phenomenon by introducing an additional keyword **phrase**. Let us consider an English to Finnish example:

Anima/Bird	eagle()	kotka
ST/Static/Property	bald(object)	kalju
Anima/Bird/Eagle	bald(Anima/Bird/Eagle)	[phrase] valkopäämerikotka

This notation should be interpreted in the following way:

- "eagle" should be translated as "kotka".
- "bald" should be translated as "kalju".
- "bald" with the dependent word "eagle" should be translated as "valkopäämerikotka", and no translation should be given for the dependent word(s).

A simple additional rule is also introduced: the first matching dictionary entry is used for the translation. The phrase "bald eagle" matches both *bald(object)* and *bald(eagle)*; in this case we select *bald(eagle)* alternative, since it is occurs earlier in the dictionary. Generally, the most specific translations of a word and compound words should be placed on top of the list of alternatives.

The mechanism described above is also used for describing idioms or slang phrases. For example:

SC/Existence	kick_1(the_1)	[phrase]
Preposition	the_1(bucket)	
Function/Container	bucket()	

4 Dictionary Representation

Examples of dictionary entries are tabulated in Tables 1 and 2. The column "ID" lists the identifier for each dictionary item. "Word" is the base form of the item. "Word Formula" defines, based on the EuroWord ontology, the concept class to which the word belongs to, and the concept classes which representatives in can link to. The variable names denoted in bold face are used to assist translation (by providing a mapping between the corresponding words), and do not refer to concept classes. Table 3 establishes connections between the words in single-language tables.

5 Conclusions

In this paper we introduced a new method for describing multilingual dictionaries in machine-readable format. In addition to multilinguality, the most interesting features of our approach are the use of concept ontologies, phrase/compound word definitions, and a formal definition of word-to-word links. While the work is still in progress, we believe that this particular approach will make the design of multilingual dictionaries more useful and accessible for semantic-rich natural language text parsers. The functionality of a dictionary of the type we propose offers more advanced and convenient features in comparison to dictionaries based on traditional approaches. Such a dictionary is able to recognize the context of the word (which leads naturally to the selection of the right translation for a particular case), and is able also to supply formal phrase patterns that help a translator to use words correctly (e.g. by using the appropriate verb government).

References

Tusov, V. A. (2004). Computer Semantics of the Russian Language (in Russian). St. Petersburg University Press, 400 p.

Mel'čuk, I. (1995). The Russian Language in the Meaning-Text Perspective. Wiener Slawistischer Almanach/Škola "Jazyki russkoj kul'tury", 682 p.

Mel'čuk, I. et al. (1999). Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV. Les Presses de l'Université de Montréal, 347 p.

[2] Illative and allative cases in Finnish can be roughly described as equivalents of English prepositions "into" and "onto". Some places require illative case ("travel into Helsinki"), while other places work with allative case ("travel onto Tampere").

Table 1. Examples of dictionary entries for English

ID	Word	Word Formula
A1	die	ST/Dynamic/Bounded(Origin/Natural/Living **who**, Preposition/Of/Cause **of-cause**)
A2	of	Preposition/Of/Cause(Situation/Component/Cause **cause**)
A3	kick	Origin/Natural/Death(Preposition/The/Bucket **bucket**) [phrase]
A4	the	Preposition/The/Bucket(Function/Container/Bucket **bucket**)
A5	bucket	Function/Container()
A6	empty	SituationComponent/Physical()
B1	kick	ST/Dynamic(Object **something**)
B2	the	Preposition/The(Anything **x**)
C1	travel	ST/BoundedEvent/Physical/Location/Move(Preposition/By/Using **by-using**)
C2	travel	ST/BoundedEvent/Physical/Location/Move(Preposition/By/Across **by-across**)
C3	travel	ST/BoundedEvent/Physical/Location/Move(Preposition/To/Place **to-place**)
C4	by	Preposition/By/Using(Function/Vehicle **vehicle**)
C5	by	Preposition/By/Across(Function/Place **place**)
C6	to	Preposition/To/Place(Function/Place **place**)
C7	car	Function/Vehicle/Object/Artifact()
C8	riverside	Function/Place/Part()
C9	Helsinki	Function/Place/Artifact()
C10	Tampere	Function/Place/Artifact()

Table 2. Examples of dictionary entries for Finnish

ID	Word	Word Formula
A1	kuolla	ST/Dynamic/Bounded(Origin/Natural/Living **who** : nominative, Cause **cause** : illative)
A2	potkaista	ST/Dynamic/Bounded(Function/Container/Tyhjä **tyhjä** : partitive) [phrase]
A4	tyhjä	SituationComponent/Physical()
A5	ämpäri	Function/Container()
B1	potkaista	ST/Dynamic/Bounded(Object **what** : partitive)
C1	matkustaa	ST/BoundedEvent/Physical/Location/Move(Function/Vehicle **vehicle** : adessive)
C2	matkustaa	ST/BoundedEvent/Physical/Location/Move(Function/Place **place** : partitive)
C3	matkustaa	ST/BoundedEvent/Physical/Location/Move(Function/Place/OnPlace **place** : allative)
C4	matkustaa	ST/BoundedEvent/Physical/Location/Move(Function/Place/InPlace **place** : illative)
C5	auto	Function/Vehicle/Object/Artifact()
C6	joenvarsi	Function/Place/Part()
C7	Helsinki	Function/Place/Artifact/InPlace()
C8	Tampere	Function/Place/Artifact/OnPlace()

Table 3. Examples of translational table entries

English	Finnish	Comment
A1	A1	die
A3	A2	kick the bucket (En) kick the emptiness (Fi)
B1	B1	kick
A5	A5	bucket
A6	A4	empty
C1	C1	travel by a vehicle
C2	C2	travel by/across something
C3	C3 / C4	travel to a place
C8	C6	Riverside
C7	C5	Car
C9	C7	Helsinki
C10	C8	Tampere

Baldwin, T., Bond, F. & Hutchinson, B. (1999). A valency dictionary architecture for machine translation. Proc. of TMI-99: 207–217.

Vossen, P. et al. (1998) The EuroWordNet Base Concepts and Top Ontology. 1998-TR-004. Centre National de la Recherche Scientifique, France.

Knight, K. & Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. Proc. of 12th National conf. on AI 1:773-778.

BMIR (2008). Protégé Ontology Library (Stanford Center for Biomedical Informatics Research). http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library (accessed: 20.05.2008).