Comparing, Integrating Lexical Definitional Knowledge From Multiple Sources

Lucja M. Iwanska

School of Computer & Information Science
Georgia Southwestern State University
800 Wheatley Street, Americus, Georgia 31709
liwanska@gsw.edu

Abstract

We discuss a computational mechanism for comparing and integrating lexical definitional knowledge, meanings and concept definitions of English words and phrases available from different sources such as dictionaries, encyclopedias, corpora of texts, and personal beliefs. Such a mechanism is needed in order to automate comparison and reconciliation of the definitional differences because completeness and correctness of definitional knowledge seriously affect the results of text processing, particularly classification, question answering, and summarization.

1 Problem Statement, Motivation

1.1 Same Word, Many Different Definitions

What is the meaning of words and phrases? What concepts do they denote? Different sources of definitional knowledge, including dictionaries, encyclopedias, various texts, and people sharing their personal beliefs, define common words such as "document" and less common such as "virus" quite differently. Their definitions differ significantly in terms of length, properties (dimensions of information), their significance, levels of specificity, the number of different senses; see Tables 1-5.

Table 1. DOCUMENT according to different sources

SourceD1	1. a piece of paper, booklet, etc., providing information, esp. of an official or legal nature. 2. qual Archaic. evidence; proof.
SourceD2	1. anything printed, written, etc., relied upon to record or prove something 2. anything serving as proof
SourceD7	1. writing that provides information (especially information of an official nature) 2. anything serving as a representation of a person's thinking by means of symbolic marks 3. a written account of ownership or obligation 4. (computer science) a computer file that contains text (and possibly formatting instructions) using 7-bit ASCII characters
SourceP1	1. something validating a claim or establishing a fact
SourceP2	1. an official-looking paper with writing and maybe a seal

Table 2. MURDER according to different sources

SourceD2	1. The unlawful and malicious or premeditated killing of one human being by another. Also, any killing done while committing some other felony, as rape or robbery. 2. Collloq. Something very hard, unsafe or disagreeable to do or to deal with.
SourceD3	1. the crime of intentionally killing a person
SourceD4	1. the crime of unlawfully killing a person especially with malice aforethought **a** something very difficult or dangerous **b** something outrageous or blameworthy
SourceD5	1. The unlawful killing of one human by another, especially with premeditated malice. 2. Slang Something that is very uncomfortable, difficult, or hazardous: 3. A flock of crows.
SourceD6	1. The offense of killing a human being with malice prepense or aforethought, express or implied; intentional and unlawful homicide.
SourceD7	1. unlawful premeditated killing of a human being by a human being
SourceP3	1. Killing someone without justifications defined by society.
SourceP4	1. The act of killing a living being is called murder. This is a crime and is against the ethics of human life.
SourceP5	1. Killing a human.
SourceA1	1. The willful (nonnegligent) killing of one human being by another.

With any information and knowledge, the reasons for differences include incompleteness and lack of knowledge, errors, lies, and misinformation, subjectivity, specific processing needs that deem certain characteristics and details as relevant and important.

Additionally, such big differences exist because it appears that natural languages are inherently ambiguous and context-dependent. Roughly, different sources give different definitions because they consider different contexts. Further complication is that words and phrases of natural language change their meanings with time. There are also regional differences.

Table 3. STUDENT according to different sources	
SourceD1	1. a person following a course of study, as in a school, college, university, etc. 2. a person who makes a thorough study of a subject 3. a person who likes to study
SourceD2	1. a person who studies or investigates 2. a person who is enrolled for study at a school, college, etc.
SourceD5	1. One who is enrolled or attends classes at a school, college, or university. 2. **a.** One who studies something. **b.** An attentive observer.
SourceD7	1. a learner who is enrolled in an educational institution 2. a learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines

1.2 Important to Know Right Definitions

This situation creates a major difficulty for designers of general-purpose natural language processing (NLP) systems. An in-depth interpretation of natural language requires a component providing lexical knowledge, a dictionary or knowledge base kind of resource. Text processing applications involving classification, summarization, or question answering may produce very different results depending on which definition will be used.

Table 4. VIRUS according to different sources	
SourceD1	1. any of a group of submicroscopic entities consisting of a single nucleic acid surrounded by a protein coat and capable of replication only within the cells of animals and plants; many are pathogenic. 2. a disease caused by a virus. 3. any corrupting or infecting influence
SourceD2	1. orig., venom, as of a snake 2. **a.** same as FILTERABLE VIRUS; specif., any of a group of ultramicroscopic or submicroscopic infective agents that cause various diseases in animals, as measles, mumps, etc., or in plants, as mosaic diseases; viruses are capable of multiplying only in connection with living cells and are regarded both as living organisms and as complex proteins sometimes involving nucleic acid, enzymes, etc. **b.** a disease caused by a virus 3. anything that corrupts or poisons the mind or character; evil or harmful influence 4. something that poisons the mind or soul 5. a computer program usually hidden within another seemingly innocuous program that produces copies of itself and inserts them into other programs and that usually performs a malicious action (as destroying data)
SourceD3	1. a very small organism, smaller than a bacterium, which causes disease in humans, animals and plants 2. Virus also means a disease caused by a virus. 3. a hidden instruction in a computer program which is intended to introduce faults into a computer system and in so doing destroy information stored in it
SourceC2	1. Viruses are extremely small infectious substances (much smaller than bacteria).

For example, the property of ‘liking to study’ and the property of ‘being enrolled at school’ have a potential to classify individuals as "students" completely differently; see definitions of "student" according to SourceD1 and SourceD5 in Table 3.

A person who understands "murder" as ‘killing a human’, see SourceP5 in Table 2, may develop a false sense of security when reading FBI statistics compiled with a different, more restrictive definition of "murder" which excludes certain types of killing a human from being classified as "murder"; FBI is SourceA1 in Table 2.

1.3 Many Competing Sources

The question arises as to which of these many definitions is the right one, the most correct and complete, and which of the many available sources should be used for building a lexical knowledge component of a NLP system, be it a dictionary or a knowledge base.

Many NLP researchers and practitioners have built and continue to build their own dictionaries/knowledge bases, which tends to be a very long and costly effort requiring serious resources. Another problem is that self-developed resources are virtually always geared toward specific applications and type of textual data processed, which contributes to the nonscalability of NLP systems.

Table 5. Sources of Information and Knowledge	
SourceD1	Collins English Language Dictionary, 1979
SourceD2	Webster's NewWorld Dictionary, 2nd College Edition, 1982
SourceD3	Cambridge International Dictionary of English
SourceD4	Merriam Webster's Collegiate Dictionary, 10th Edition
SourceD5	American Heritage Dictionary of the English Language, 4th Edition, 2000
SourceD6	Online Plain Text English Dictionary Kyoto Notre Dame University, Project Gutenberg Etext of Webster's Unabridged Dictionary
SourceD7	WordNet 1.7 Princeton University, Cognitive Science
SourceE1	Encyclopedia.com updated Columbia Encyclopedia, 6th Edition
SourceP	Personal beliefs, knowledge of different individuals
SourceC1	Knowledge automatically acquired by our NLP system from a corpora of texts
SourceA1	FBI Uniform Crime Reporting: Data Collection Guidelines

Many researchers utilize existing sources. WordNet (Fellbaum, 1998) is a wonderful and free-of-charge resource designed specifically for the needs of computational linguistics (CL) community and the dictionary of choice for many NLP systems (Voorhees and Buckland, 2002). It is not, however, the only, the best, or the most comprehensive source. There are hundreds of other sources of lexical definitional knowledge available at, among others, OneLook.com and YourDictionary.com Dictionary Search websites.

A promising recent approach pursued by a number of NLP and CL researchers is developing knowledge acquisition and learning methods to automatically create dictionaries and knowledge bases or augment the existing ones with system-acquired knowledge from corpora of texts (Iwanska et al., 1999, 2000a), (Harabagiu and Moldovan, 2000), (Rapaport and Kibby, 2002), (Reiter and Robertson,2003), (Thompson and Mooney, 2003).

1.4 Need for Comparison, Integration

Given the variety of sources and definitions for virtually all words and phrases, a comparison mechanism is needed in order to address the question as to which of the sources is the best, the most complete and correct, which definition(s) to use, and, if multiple definitions are valid, in order to identify their similarities and differences.

We developed a computational mechanism to automatically compare and, in some cases, integrate knowledge from multiple sources. Given two definitions of a word or phrase, our system computes quantitative measure of distance between them based on qualitative relations between these definitions: PARTIAL-OVERLAP, MORE-SPECIFIC / MORE-GENERAL, DISJOINT. It highlights similarities and differences. Computed comparison is used to reach the integrate-or-not decision. If integration is deemed appropriate, the system computes integrated definitions.

In our NLP system, we address incompleteness and changes in meaning through integration of our hand-crafted, modest size dictionary with definitions from reliable sources. Our primary sources include existing "respectable" dictionaries, see Table 5, and knowledge acquired automatically by our system from corpora of "respectable" texts.

Automatic knowledge acquisition methods are particularly useful for acquiring and updating phrasal definitional knowledge. For example, none of the above mentioned hundreds dictionaries define phrases such as "safe environment" or "very fast actions", both of which were learned by our system (Iwanska et al., 1999, 2000a).

Additionally, knowledge acquired from recent texts allows our system to update definitions that changed with time. For example, the fourth definition of "document" given by SourceD7 (WordNet), probably about ten years ago, is now too restrictive. Currently, any character, not just 7-bit ASCII character, can be used in a document. Knowledge acquired by our system allowed us to correctly generalize this definition to account for this change.

The capability of comparing and integrating lexical knowledge results in improved performance of our NLP system. For example. In question answering, new questions can be answered, correctness of some answers is improved, and some questions can be answered more completely. In tasks involving classification, groupings arrived via different definitions may be compared and predicted.

The rest of the paper is organized as follows: Sect. 2 provides a high-level discussion of our meaning and knowledge-level representation of text; Sect. 3 gives algorithmic details of our comparison and integration approach; it also provides a number of examples; Sect. 4 and 5 discuss reliable and unreliable sources and more details about our integration mechanism.

2 NL-Motivated Representation of Text

We discuss briefly our natural language-motivated representation of text. Further details, including question answering, representation and reasoning with text conveying spatio-temporal and probabilistic information and knowledge can be found in (Iwanska, 1993), (Iwanska, 1996), (Iwanska, 2000b).

2.1 Text as Sets of Type Equations

We represent text by natural language-motivated type equations with Boolean, set and interval-theoretic semantics of the following form

$$P == P_1, P_2, ..., P_N .$$

where P's are properties corresponding to text fragments such as noun phrases and verb phrases. Each property is a term, a record-like, graph-like, underspecified structure that consist of two elements

1. head, a type symbol, and
2. body, a possibly empty list of attribute-value pairs

attribute => value where attributes are symbols

and values are single terms or sets of terms.

For example, the sentence "Viruses are extremely small infectious substances" is represented by the equation

```
virus ==
substance(size   => small(degree => extremely),
          infect => infectious) .
```

whose right handside contains one property, a term with "substance" as its head and two attributes:

1. the attribute "size" with the value

   small(degree => extremely) which itself is a term with the type "small" as its head, and one attribute "degree" with the value "extremely".

2. the attribute "infect" with the value "infectious" which is a basic type.

2.2 Boolean, Set and Interval-Theoretic Semantics Motivated by Natural Language

Semantically, terms are subtypes of their head types. For example, the above term represents this subset of things of the type "substance" for which the attribute "size" has the value "extremely small" and for which the function "infect" yields the value "infectious".

The Boolean operations of MEET, JOIN, and COMPLEMENT simulate conjunction, disjunction and negation in natural language. They take terms as arguments and compute conjunctive, disjunctive, and complementary terms with the set-intersection, set-union, and set-complement semantics.

Efficient computation of arbitrary Boolean expressions allows the system to compute a number of semantics relations among terms, including EQUAL reflecting set identity, ENTAILMENT (and SUBSUMPTION, its dual) reflecting set-inclusion, PARTIAL-OVERLAP, reflecting non-empty set-intersection, DISJOINT reflecting empty set-intersection. These relations allow the system to compute consequences of knowledge expressed by text, and therefore compute answers to questions of the knowledge base created as the result of processing input texts, and to update system's knowledge base.

Knowledge bases with such type equations are used bi-directionally: for answering questions about the properties of entities and concepts in the left handsides, and for matching particular properties against the right handside properties of entities and concepts that the system knows about. We use these capabilities to compute comparison as well as integration of properties in different concept definitions.

3 Algorithmic Details

3.1 Input

1. Concept C, a word or phrase. For example, we may be concerned with the meaning (concept definition) of the word "virus" or the phrase "very fast actions".

2. Two knowledge sources Source1 and Source2. Our sources of definitional knowledge include dictionaries, encyclopedias, personal beliefs obtained via knowledge engineering methods, and knowledge automatically acquired by our NLP system from corpora of texts; see Table 5.

3. Concept definitions according to both sources

Source1: $\{T_{1,1}, T_{1,2}, ..., T_{1,N}\}$
Source2: $\{T_{2,1}, T_{2,2}, ..., T_{2,M}\}$

where each definition $T_{i,j}$ is text, some number of sentences or phrases such as noun phrases or verb phrases. For example, if the word is "virus" and we consider SourceD1 as Source1, and SourceD2 as Source2, then N=3 and M=6, i.e., we have three definitions of "virus" from Source1 $\{ T_{1,1}, T_{1,2}, T_{1,3}\}$ and competing six definitions of "virus" from Source2 $\{ T_{2,1}, T_{2,2}, T_{2,3}, T_{2,4}, T_{2,5}, T_{2,6} \}$. These definitions correspond to different senses; note that SourceD2 distinguishes two senses 2a., 2b.; see Tables 4 and 5.

3.2 Steps

Step 1 Compute representations of word or phrase C and of each of its textual definitions $T_{i,j}$.

Step 2 For each pair $(T_{1,k}, T_{2,n})$ of definitions from both sources, compute qualitative relation R between each pair of properties $(P_i^{1,k}, P_j^{2,n})$ in the right handside of the definitions;

R can be one and only one of the following: EQUAL, SMALLER (MORE-SPECIFIC), LARGER (MORE GENERAL), PARTIAL-OVERLAP, or DISJOINT.

Step 3 For each pair $(T_{1,k}, T_{2,n})$ of definitions from both sources, compute numeric measure of closeness D between two definitions.

This measure whose motivation is similar to (Resnik 1999) is a number between 0 and 1 computed based on qualitative relations R among the properties in both definitions and on proportion of relations indicating closeness; EQUAL corresponds to 1, the smallest distance, SMALLER and LARGER to 0.8, PARTIAL-

OVERLAP to 0.6, and DISJOINT to 0, the largest distance.

Step 4 Compute alignment of definitions based on metric D computed for each pair. This alignment shows which definitions from both sources resemble each other most closely. For the definitions of "virus" according to SourceD1 and SourceD2, see Table 4, this alignment is

((1, 2a), (2, 2b), (3, 3), (-, 1), (-, 4), (-, 5)).

Step 5 For each pair of aligned definitions, decide if integrate and choose integration mode based on the reliability of sources and on the value of D.

Step 5a Compute integrated definition. This integration, illustrated by examples in Sections 4 and 5, involves computing the Boolean operations of meet (conjunction), join (disjunction), and complement (negation) on the properties in the right handside of the definitions.

Step 5b Generate English text for the integrateddefinition.

Step 5c Update system dictionary/knowledgebase with the integrated definition.

3.3 Output

1. Updated system dictionary/knowledgebase incorporating knowledge from both sources.

2. Alignment of definitions

3. Highlights of similarities and differences between pairs of definitions.

4 Reliable and Unreliable Sources

Depending whether sources are reliable or not (in general or in terms of specific piece of information or knowledge), we use different integration operations. If both are reliable, we integrate most aggressively and the resulting integrated piece reflects fully all that both sources provided. If one source may not be reliable, a conservative integration is performed. Finally, if a source is known or suspected to be unreliable, we first negate its information and then fully combine it with all provided by the reliable source.

Consider temporal information about the occurence of an event provided by two sources, different people recalling the same event.

Source1: "It took place in 1992, April or May"
Source2: "It did not happen in early May"
Depending whether these sources are considered reliable or not, we combine their information differently, which results in three possible integrated information about the time when the event took place. Information provided by the sources is translated into the following terms

D_1 = date(month => [April, May],
year => 1992)

D_2 = date(month => not May(part => early)) =
date(month => [not May, May(part => not early)])

4.1 Both Sources Reliable

If both sources are considered reliable, we use the meet operation to compute integrated piece of information or knowledge. This operation, a conjunction with inheritance, incorporates fully all information provided by both sources. For the above dates, an integrated term is computed

D_1 MEET D_2 = D^1 =

date(month => [April, May(part => not early)],
year => 1992)

which gets automatically translated into an English phrase "April or May, but not early May, 1992".

4.2 One Source Possibly Unreliable

If one source may not be reliable, but it is not known which one, we use the join operation to integrate. This operation, a disjunction, incorporates conservatively information provided by both sources. For the above dates, the integrated term cannot be simplified, its two elements are partially overlapping because both sources provide different aspects of the temporal information.

D_1 JOIN D_2 = D^2 =

[date(month => [April, May],
year => 1992),
date(month => not May(part => early))]

which gets automatically translated into a disjunctive English phrase "April or May, 1992, or not early May".

4.3 One Source Reliable, One Unreliable

If one source is considered unreliable, eg. it is known or suspected to have lied or to be ignorant, we use the complement operation to negate its information. The rationale is that if information or piece of knowledge

is incorrect, then the actual correct information and knowledge, whatever it may be, is consistent with the negation of what the source provided. The complement operation allows us to capture this. We then integrate both terms via the meet operation. For the above dates, the system computes an integrated term

D_2^{neg} = not D_2 = [not date, date(month => May(part => early))]

D_1 MEET D_2^{neg} = D^3 =
date(month => May(part => early),
year => 1992)

5 Reliable Sources

5.1 Partially Overlapping Concepts

Definitions from different sources frequently denote partially overlapping concepts. Overlap exists because properties are described at different levels of specificity and because some properties are stated only by one source. If both sources are reliable, we mostly use the most aggressive mode of integration, which combines all knowledge provided by both sources. In the integrated definition, some properties become more specialized (more informative) and some other new properties are added.

An example is a dictionary definition which we update with knowledge acquired from texts. As shown in Table 4, SourceD3 defines "virus" as "a very small organism, smaller than a bacterium, which causes disease in humans, animals and plants", and SourceC1 as "extremely small infectious substances (much smaller than bacteria)".

The integration of the first definition with the second produces an integrated definition "an extremely small, infectious organism (substance), much smaller than a bacterium, which causes disease in humans, animals and plants". Two size-related properties get more specialized: "very small" becomes "extremely small", and "smaller" becomes "much smaller". These integrated properties contain strictly more information than (entail) the corresponding properties in the old definition. The new property added is "infectious". This is accomplished as follows.

First, the representation of definitions is computed

$virus_{(SourceD1)}$ == $P_1^{1,1}$, $P_2^{1,1}$, $P_3^{1,1}$.
$virus_{(SourceC1)}$ == $P_1^{2,1}$, $P_2^{2,1}$.

$P_1^{1,1}$ = organism(size => small(degree => very))

$P_2^{1,1}$ = smaller(arg2 => bacterium)

$P_3^{1,1}$ = causes(np => disease(pp => in(np => [humans, animals, plants])))

$P_1^{2,1}$ = substance(size => small(degree => extremely),
infect => infectious)
$P_2^{2,1}$ = smaller(quantity => much,
arg2 => bacterium) .

Then, relations R for each pair of properties in the right handside of the equations are computed via the meet operation.

R($P_1^{1,1}$, $P_1^{2,1}$) = PARTIAL-OVERLAP because P_1 = $P_1^{1,1}$ MEET $P_1^{2,1}$ =
organism(size => small(degree => extremely),
infect => infectious)

P_1 LESS-THAN $P_1^{1,1}$ and P_1 LESS-THAN $P_1^{2,1}$

R($P_2^{1,1}$, $P_2^{2,1}$) = MORE-GENERAL (LARGER) because P_2 = $P_2^{1,1}$ MEET $P_2^{2,1}$
=
smaller(quantity => much,
arg2 => bacterium)

P_2 LESS-THAN $P_2^{1,1}$, P_2 = $P_2^{2,1}$

The relations R for the other pairs of properties are DISJOINT because the meet operation yields terms corresponding to empty set. D = 2/3 and in the COMBINE-ALL integration mode, the integrated type equation has three properties: the integrated properties P_1 and P_2, and the unchanged property $P_3^{1,1}$. The integrated equation is

virus ==

organism(size => small(degree => extremely),
infect => infectious) ,

smaller(quantity => much,
arg2 => bacterium) ,

causes(np => disease(pp => in(np => [humans, animals, plants]))) .

This equation then gets translated into English phrase "an extremely small, infectious organism (substance), much smaller than a bacterium, which causes disease in humans, animals and plants", in which the order of properties mentioned follows the order in the original definition.

5.2 Concepts in MORE-GENERAL Relation

Definitions from different sources may denote concepts in MORE-GENERAL (LARGER) relation. For example, as the following equations reveal, SourceP3 definition is strictly more general, i.e., denotes larger set, than definitions from SourceD3 and SourceA1.

$murder_{SourceD3}$ == killing(intent => intentionally,
object => person) .

$murder_{SourceP3}$ == killing(object => human) .

$murder_{SourceA1}$ == killing(intent => wilful,
agent => human,
object => human) .

Such a relation may indicate that one source has a definition that is too general due to, for example, ignorance. It can also indicate that a source has a definition that is overly specific, i.e., not generalized enough. We do not have means to automatically decide which is the case. In certain cases, we make somewhat arbitrary assumptions.

For example, if two dictionary definitions are in MORE-GENERAL relation, we integrate by keeping the most specific. Then, if context requires certain properties at given level of specificity, we generate shorter, more general definitions via our summarization/generalization mechanism. In case of personal beliefs, unless a person is known to be an expert, we assume that sources such as dictionaries and texts are more correct and integrate accordingly.

5.3 Clashes Signal Need to Generalize, Correct

Clashes between information and knowledge from different sources indicate inconsistencies that need to be resolved. In our representation, inconsistencies are automatically detected when the meet operation generates a term corresponding to empty set.

Some clashes indicate the need to generalize, others reflect errors or deliberate misrepresentations that need to be corrected. We have a mechansim to identify clashes, but we do not have automatic way to decide what to do about them. Each time the system generates a clash, a human has to make the decision what to do about the clash.

This situation is reminiscent of expert systems and knowledge-based systems in that the decision which piece of knowledge or which expert is correct does not appear to have a general solution and involves rather arbitrary assumptions and trust.

5.4 Integrating Two Word Senses Into One

The same source, eg. dictionary, can be used as if two sources, which allows us to investigate similarities and differences between different senses of the same word or phrase. In some cases, similarities lead to integrating two senses into one, thus reducing the number of word senses.

For example, similarity between partially overlapping senses of “virus” , see Table 4 for definitions 3 and 4 from SourceD2, led to one combined sense. The original two senses “anything that corrupts or poisons the mind or character” and “something that poisons the mind or soul” are represented as follows

virus == [corrupt, poison](object => [mind, character]) .

virus == poison](object => [mind, soul]) .

The conservative, join operation integration combined with a machine learning-style inductive leap (add or skip some aspect in order to simplify and/or shorten the utterance) results in one combined word sense which corresponds to the first original sense.

6 Discussion, Ongoing and Future Efforts

6.1 Short versus LongTexts

Our integration mechanism appears to work well when textual definitions are short texts with not very long sentences and phrases. This is the case with standard dictionaries and our system acquired knowledge which, by design, acquires knowledge in small portions, short, few sentence-long texts. We can accomplish this because for short texts, parsing and computing meaning-level representation is possible and can be done with high levels of precision.

Full integration of larger texts such as many page encyclopedic entries or complete newspaper articles is currently not really possible because parsing long sentences and computing meaning-level representation of large texts with high levels of precision remains an open research problem.

6.2 Integrate or Not

A really hard part is the integrate-or-not decision. In general, it is hard both for humans and systems to decide who is right and which piece of knowledge is correct. So despite having a system capable of fully automatic integration, we involve a human in the loop. We look at the system's recommendation, the alignment of different definitions, similarity metrics etc. and then make this decision by hand.

The only alternative appears to make an arbitrary assumption that particular sources are (always) right or more right than some others.

We have a mechanism that, in principle, allows us to integrate definitions from all existing sources. In practise, we consider a safer road of choosing two existing sources and updating them only with knowledeg acquired automatically by our system from corpora of “respectable” texts.

6.3 Dictionary Entries as Summaries, Generalizations

Our investigation leads us to believe that dictionary entries may be summaries and generalizations of words' uses over certain contexts. As such, they would constitute derived, and not primary, resource in people and machines. We plan to continue developing knowledge acquisition and learning methods to automatically create dictionaries/knowledge bases from corpora of texts. Our approach is to let the system acquire as much as possible and as specific as possible pieces of information and knowledge. We then generate dictionary-like, short, context-relevant definitions via our summarization/generalization mechanism.

6.4 Text Generation

Even with shorter text, we encounter many problems with generating naturally-sounding English text from our representation. One problem is that integration results in increasingly heavier phrases. Breaking long phrases into separate sentences with shorter phrases sometimes produces akward texts.

Another problem is naturalness, which may mean different things in different contexts. In case of synonymous relations, we use two criteria. The first is frequency, commonality-based with preference given to the more commonly used, relative to a corpus, subject matter, or overall. For example, the word “infectious” will be preferred to “pathogenic”, and the phrase “extremely small” to “submicroscopic”. The second criterion is based on simplicity and size of utterance. For example, the word “submicroscopic” will be preferred to “extremely small”.

It is clear that with progress on processing larger texts, the text generation problems will intensify.

References

Fellbaum, C., ed 1998. *WordNet An electronic lexical database* , The MIT Press.

Harabagiu, S. and Moldovan, D. 2000. Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text. in Iwanska, L.M., and Shapiro, S.C. eds 2000. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language* The MIT Press. 301-334.

Iwanska, L. 1993. Logical Reasoning in Natural Language: It Is All About Knowledge. *International Journal of Minds and Machines,* **3**(4): 475-510.

Iwanska, L. 1996. Natural (Language) Temporal Logic: Reasoning about Absolute and Relative Time". *International Journal of Expert Systems,* **9**(1): 113-149.

Iwanska, L., Mata, N. and Kruger, K. 1999, 2000a. Fully Automatic Acquisition of Taxonomic Knowledge from Large Corpora of Texts: Limited-Syntax Knowledge Representation System based on Natural Language. *Proceedings of the Eleventh International Symposium on Methodologies for Intelligent Information Systems (ISMIS99)*, Springer-Verlag, pp. 691-697, Warsaw, Poland, 1999. Reprinted in Iwanska, L.M., and Shapiro, S.C. eds 2000. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language* The MIT Press. 335-346.

Iwanska, L. 2000b. Natural Language is a Powerful Knowledge Representation System: The UNO Model. In Iwanska, L.M., and Shapiro, S.C. eds 2000. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language* The MIT Press, 7-64.

Liu, H. 2003. Unpacking Meaning from Words: A Context-Centered Approach to Computational Lexicon Design in Blackburn, P. et al. eds 2003 *Modeling and Using Context* , Proceedings of the 4th International and Interdisciplinary Conference, Lecture Notes in Artificial Intelligence 2680, Berlin, Springer, 218-232.

Muresan and Klavans, 2002. Muresan, S. and Klavans, J. A method for automatically building and evaluating dictionary resources. *Proceedings of the Language Resources and Evaluation Conference (LREC 2002).*

OneLook.com Dictionary Search

Rapaport, W.J. and Kibby, M.W. 2002. Contextual Vocabulary Acquisition: A Computational Theory and Educational Curriculum. in *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)*, Orlando, FL, 2002.

Reiter, E. and Robertson. R. 2003 Acquiring Correct Knowledge for Natural Language Generation *Journal of Artificial Intelligence Research* 18:491-516.

Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11:95-130.

Thompson, C.A. and Mooney, R.J. 2003. Acquiring Word-Meaning Mappings for Natural Language Interfaces *Journal of Artificial Intelligence Research* 18:1-44.

Voorhees, E.M. and Buckland, L.P. eds 2002. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002).* Department of Commerce, National Institute of Standards and Technology.

YourDictionary.com Dictionary Search