Journal of AI and Data Mining
Vol 7, No 1, 2019, 109-119
DOI: 10.22044/JADM.2018.4966.1601

Automatic Construction of Persian ICT WordNet using Princeton WordNet

A. Ahmadi Tameh, M. Nassiri* and M. Mansoorizadeh

Computer Department, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran.

Received 08 November 2016; Revised 14 August 2017; Accepted 05 April 2018
**Corresponding author: m.nassiri@basu.ac.ir (M.Nassiri).*

Abstract
WordNet is a large lexical database of the English language in which nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets). Each synset expresses a distinct concept. Synsets are inter-linked by both semantic and lexical relations. WordNet is essentially used for word sense disambiguation, information retrieval, and text translation. In this paper, we propose several automatic methods to extract Information and Communication Technology (ICT)-related data from Princeton WordNet. We then add these extracted data to our Persian WordNet. The advantage of automated methods is to reduce the interference of human factors and accelerate the development of our bilingual ICT WordNet.
In our first proposed method, based on a small subset of ICT words, we use the definition of each synset to decide whether that synset is ICT. The second mechanism is to extract the synsets that are in a semantic relation with the ICT synsets. We also use two similarity criteria, namely *LCS* and S^3M, to measure the similarity between a synset definition in WordNet and definition of any word in Microsoft dictionary. Our last method is to verify the coordinate of ICT synsets. The results obtained show that our proposed mechanisms are able to extract the ICT data from Princeton WordNet at a good level of accuracy.

Keywords: *WordNet; Semantic Relation; synset; Part of Speech; Information and Communication Technology.*

1. Introduction

Semantic Network is one of the famous structured tools for data representation. A well-known example of such semantic network is the work of George Miller and his colleagues accomplished in Princeton University ([1], [2]). They developed a lexical semantic network concept and constructed WordNet.
WordNet is a lexical database that groups synonymous words into the so-called synsets, and relates the synsets with various semantic information like *hypernymy*, and *meronymy*. For each synset, it also provides short definitions and usage examples. WordNets have been extensively used in computer processing of natural languages and applications such as word sense disambiguation, information retrieval, text classification and summarization, and machine translation. Princeton WordNet prepared the road for constructing other WordNets. For instance, Greek WordNet in Computer Science and Psychology [3] is a specialized WordNet.
Recently, computer processing of Persian language has gained much interest in the Iranian academic and research community ([4], [5]). A lexical resource like WordNet has a key role in software applications such as semantically-enriched Persian search engines, text classification, and machine translation. There have been several efforts towards developing Persian or bilingual Persian-English WordNets ([6], [7]). However, the existing Persian WordNets are limited in both word and semantic relations coverage. More precisely, they only cover the general domain of the language, and they only record a subset of semantic relation types. These limitations have encouraged researchers to develop specialized WordNets with the aim of expanding WordNet in terms of both words

and relation types. In this regard, we started to Information and Communication Technology (ICT) domain. We have undertaken a variety of resources and approaches to select the ICT terms, define semantic relation types and relations, and translate the terms to Persian.

In this work, we report a part of this process, in which we try to extract ICT information from the existing WordNets. The final product is a bilingual ICT WordNet, which is targeted to be used by public users as well as researchers. In this paper, we aim at extracting the data belonging to the ICT domain from the Princeton WordNet. We then integrate the extracted data in our bilingual Persian-English WordNet of the ICT domain.

We define ICT as a collection of devices, tools, and methods for generating, processing, transmitting, and manipulating information. According to this definition, radio, television, fax, computer and Internet, printer, scanner, and digital camera are all examples of the ICT concepts. Software and algorithms for data storage and data analysis are also categorized in the ICT domain. As more examples, different types of communication networks and protocols that are used for handling data transmission belong to ICT.

Our strategy is to first develop an ICT WordNet in the English language, as we have access to a huge English content in this domain via web and offline data stores. In the second step, we translate our ICT WordNet into the Persian language.

By wide spreading the usage of ICT words in the society, the Princeton WordNet is increasingly incorporating such concepts along with their corresponding semantic relations. On the other hand, this subset covers the basic concepts of the Internet, networking, data processing and communications, and the whole information and communication technology domain. This accurate and useful data can be regarded as the core of ICT WordNet. This initial network can be further enriched by extending super-concepts and including their various sub-classes. For instance, the *network protocol* is in Princeton WordNet. However, specific protocols for different tasks in network communications are not included. Therefore, it is easily possible to extend the semantic network around the *network protocol*.

While the task seems to be trivial at the first glance, it is quite challenging as some of the main ICT terms have several meanings and usages in other domains. As an example, *communications* has a wide usage in social domain and humanities. This intrinsic construct a bilingual WordNet for the ambiguity can be tackled by using sense disambiguation techniques such as word co-occurrence analysis in text corpora. The results obtained by our proposed algorithms are comparable to that of the state of the art.

Note that the extracted data from Princeton covers only a small part (about 8%) of our bilingual WordNet. As mentioned earlier, the remaining part of our WordNet is constructed using text mining methods applied to a big ICT corpus. However, in this paper, we focus on our propositions to extract the ICT data from Princeton.

In this paper, we propose several techniques to extract the ICT-related data from Princeton WordNet. In our first proposed method, based on a small subset of ICT words, we use the definition of each synset to decide whether that synset is ICT. The second mechanism is to extract the synsets that are in a semantic relation with the ICT synsets. We also use two similarity criteria, namely Longest Common Subsequence (*LCS*) and Sequence and Set Similarity Measure (S^3M), to measure the similarity between a synset definition in WordNet and definition of any word in Microsoft dictionary. Our last method is to verify the coordinate of the ICT synsets.

The rest of this paper is organized as follows. We review the state of the art in Section 2. In Section 3, we propose several mechanisms to extract the ICT-related data from Princeton WordNet. In Section 4, we discuss the results obtained by each method. Finally, in Section 5, we conclude the paper.

2. Related works

In this section, we review the main features of Princeton WordNet as well as some other general and specialized WordNets in different languages. A general WordNet covers all the domains in a language. Therefore, a deep knowledge of the language is required for its construction. However, a specialized WordNet covers words in a specific domain like agriculture, medicine, and computer science. Here, in addition to language knowledge, expertise in that domain is also required.

2.1. Semantic network of words

WordNet is an enhanced dictionary in which words are classified based upon their meanings. In WordNet, the synonym words are grouped in a structure called *synset*. Each synset represents a separate concept. However, there are several differences between a WordNet and a dictionary. First, a WordNet not only

connects the lexical parts of words but also connects their concepts. Therefore, the words located close to each other in the WordNet also have a semantic proximity. Secondly, WordNet tags the semantic relation between words, while classification of words in a classical dictionary is only based upon their lexical similarity, and it does not specify any semantic relation between words.

Synsets are inter-linked to each other using semantic as well as lexical relations. Some well-known semantic relations in the WordNet are *hypernym/hyponym*, *meronym/holonym*, and *domain* relationships.

There are several ways to construct a WordNet: manual, semi-automatic, and automatic. In the manual method, human experts manually add new data including words, synsets, and semantic relations to the WordNet database. In automatic methods, the WordNet data is extracted using text-mining approaches, and the intervention of human experts is minimized. A semi-automatic method is a hybrid approach that uses both the manual and automatic mechanisms. The manual construction of WordNet is time-consuming and error-prone. Furthermore, it requires a high level of language knowledge. However, automatic construction decreases the involvement of the human factor and increases the construction speed.

A product like WordNet is mainly used for word-sense disambiguation (WSD), information retrieval, and text translation. In order to speed-up the development process, we decided to deploy the automatic approaches for constructing ICT WordNet. We believe that there are two possible ways to automatically construct Persian WordNet in the ICT domain. In the first method, ICT WordNet is first constructed in the English language and then translated to Persian, while in the second method, using a bilingual dictionary and Princeton WordNet, the ICT concepts are extracted from English WordNet and then added to Persian WordNet. In this work, we selected the first method to construct our bilingual ICT WordNet.

Several general WordNets such as Arabic [8], Russian [9-10], Japanese [11], French [12], and Swedish [13] have been constructed by mapping and translating the well-known Princeton WordNet. However, constructing specialized WordNets such as Greek WordNet in the psychology domain, in addition to the global language knowledge, requires expertise in psychology. In other words, more skills and expertise are vital for constructing a specialized WordNet due to the need to the expert linguist and wide variety of semantic relations compared to a general WordNet.

2.2. General WordNets

Princeton WordNet mainly covers concepts in general language domain. In addition to Princeton WordNet, general WordNets in different languages like Persian, Korean, Swedish, Japanese, Arabic, Russian and French have been developed [14]. In the following subsections, we briefly review some of these WordNets.

2.2.1. Princeton WordNet

Princeton WordNet was the pioneer work in this domain. It is now in its 3.1th version. In Princeton WordNet, information is organized based on a logical group named as *synset*. Each synset includes a set of synonym words and pointers that explain the relations between this synset and other synsets. Words in one synset are classified in a way that they could be replaceable in some texts. It is also likely that one word or collocation appears in more than one synset.

The semantic and lexical relations are two kinds of relations that are demonstrated by pointers. The lexical relations are established between *forms* of words that are semantically related. However, a semantic relation intensely helps understanding the correct meaning of a word and its application in logical deduction. There are almost 30 semantic relations between synsets in Princeton WordNet, e.g. *hypernymy/hyponymy*, *meronymy/holonymy*, *implications*, *cause*, *and similarity*.

WordNet 3.1 covers four parts of speech (POS) including nouns, verbs, adverbs, and adjectives. Approximately 70% out of 117659 synsets in Princeton are nouns, 15% are adjectives, 12% are verbs, and almost 3% are adverbs. Princeton WordNet was constructed manually, and has been a basis for constructing many other WordNets.

2.2.2. FarsNet: a Persian WordNet

FarsNet [7] is a WordNet in the Persian language constructed based on Princeton WordNet 2.1. The current data in this WordNet is the result of several automatic extraction techniques applied to different Persian corpora. They use two Persian and English corpora as well as a bilingual dictionary in order to map English synsets of Princeton into Persian synsets in FarsNet. Each Persian word could have several English translations and each English translation

could also belong to several Princeton synsets. Hence, in the first step, for a specific Persian word, a bilingual dictionary is used to extract equivalent English words. Then a set of synsets is selected with the help of Princeton WordNet, which includes all English translations of this Persian word.
According to this mapping, if the English translation of a Persian word has only one sense in Princeton WordNet, the Persian synset is directly linked to the corresponding synset in Princeton WordNet. On the other hand, if two or more English translations exist, a matching score is computed for each synset and a synset with the highest score is selected as an appropriate synset for that word.

2.2.3. Korean WordNet

Korean WordNet is automatically constructed by means of several disambiguation techniques for connecting Korean words to the synsets in Princeton through a bilingual dictionary [15]. The main issue is that when Korean words are linked to the WordNet synset, it may cause a semantic ambiguity. In order to overcome this problem, several heuristic solutions related to sense disambiguation are deployed in constructing Korean WordNet. These solutions exploit various concepts such as the maximum similarity between a Korean word and the English synset, *IS-A* relations between an English word and its corresponding Korean word, etc. Finally, these solutions are combined to determine whether the synset can be linked to the word or not. They use a decision-tree to combine the solutions. A manually provided training set is used for tree induction. The tree decides if a Korean word could be linked to an English synset or it should be thrown away.

2.3. WordNet expansions

There exist some works that expand WordNet coverage and combine it with other knowledge sources. Here, we briefly review some notable cases since they try to automatically match and relate the WordNet synsets to external words or documents.
BabelNet [16] is a multilingual wide coverage semantic network that integrates six public domain lexical resources. Its main goal was word sense disambiguation. BabelNet maps Wikipedia pages to the most similar WordNet synsets according to their headword. The mapping takes into account word contexts as well as Wikipedia page redirections.
Nimb and Pedersen map a Danish thesaurus to Danish WordNet to get a more structured representation of the knowledge available in thesaurus [17]. The initial results are promising with over 90% success in a correct matching of relationships.
Several recent efforts address the problem of shortcoming of Princeton WordNet in covering new words, especially words and phrases developed in social media and everyday language usage. CROWN [18] is an extension of WordNet that adds novel words and phrases from Wiktionary[1] to the Princeton WordNet. It is generated automatically and includes the everyday words, social media terms, and slang.
Colloquial WordNet is another effort to extend WordNet by including the English terms and phrases used in everyday communications in social media [19]. As opposed to CROWN, this resource is developed manually by human annotators and uses Twitter and Reddit as the main sources.
In his PhD thesis, Johnatan Rusert extends WordNet by automatically adding technical terms and social media [20]. The approach adopts word embedding, and tries to find the best place for the new word in the semantic hierarchies. While this approach is very efficient in terms of finding proper location, it requires a short definition for new words.

2.4. Specialized WordNets

Applications of specialized WordNets are more or less the same as those for general WordNets. However, constructing a specialized WordNet requires a broad knowledge in that domain. Heretofore, very few specialized WordNets have been constructed. In what follows, we briefly introduce two of these WordNets.

2.4.1. WordNet domain

WordNet Domains [21] is a project in which synsets in Princeton WordNet (version 2.1) are classified according to their subject domains. Each synset is assigned with a label representing its domain. Since a word may belong to multiple synsets, this word may be assigned with several labels. Before using these labels, it is necessary to map the WordNet 2.1 synsets to the WordNet 3.0 ones. This mapping has already been published by the NLP group in Cataluna University of Spain. These mappings are not one-to-one, and some synsets in version 2.1 could be mapped into two or more synsets of version 3. Here, we assign the same domain label to all synsets in version 3 that corresponds to a specific synset in

[1] https://www.wiktionary.org/

version 2.1. Table 1 shows several labels in WordNet Domains.

Table 1. Sample labels for synsets in Domain WordNet.

Domain Samples
Film | motion picture | motion-picture show | movie | moving picture | moving-picture show | pic | picture | picture show → *Racing, sociology, telecommunication*
Computer | computingdevice | computingmachine | dataprocessor | electroniccomputer | information processing system → *computer science*
cat|computed axial tomography | computedtomography| computerized axial tomography| computerizedtomography | CT → *computer science, radiology*
color television tube | colortube | colortvtube | colour television tube | colourtube | colourtv tube → *computer_science, radiology*

2.4.2. German Bio-WordNet

The objective of Bio-WordNet construction was to cover words in the biology domain but during the preliminary steps, it was concluded that English WordNet was inadequate for covering the required semantic relations. The authors claimed that it was due to the incompatibility between the English WordNet designs and the lexical relations between words in the biology domain. Hence, the project was stopped [22].

2.4.3. Greek Wordnet expansion to psychology and computer science domains

Kremizis et al. expanded Greek WordNet to cover the psychology and computer science domains [3]. Their report shows that their mechanisms to extract specialized words and phrases in psychology and computer science are similar to those deployed in general WordNet. Some words with several meanings in general domain may also have special meanings in a specific domain. The authors maintained two copies of a word to clarify its general and special senses.

Finally, WordNet is now considered by the research community as a rich and interesting language resource. Many researchers have tried to extend it to other languages and/or specialized domains. However, this task has its own non-trivial challenges. First, the relations and structures defined in WordNet are very static and tightly coupled to the implementation of the database files. Minor modifications such as adding a new type of semantic relation or changing multiplicity of the existing relations are very hard to apply or even impossible with the current structure of WordNet. Secondly, although WordNet was mainly proposed to be used by computers, it always inherits parts of intrinsic ambiguity in natural languages. For instance, many parent-child relationships and incoherencies exist in WordNet that even the linguists do not agree upon. Unnecessary details and levels in the hierarchies are difficult to follow. As an example, ***apple*** and ***fruit*** are far apart in their semantic hierarchy, while ***keyboard*** and ***device*** are directly related. As a result, implementing this long and strictly-defined chain of concepts is not trivial, specifically in specialized WordNets.

We believe that any new attempt to build a WordNet for a specific domain should only take the Princeton WordNet as an inspiration. More precisely, one may adopt more flexible software technologies. For example, new types of relations, specifically the semantic ones, should be defined in accordance with the desired domain.

3. Construction of ICT WordNet

The most accurate approach available to extract the ICT synsets from WordNet is that an expert manually inspects each single synset. However, this brute-force and subjective procedure requires an extensive human effort, which is very costly. Furthermore, mitigating the effect of subjectivity requires subsequent revisions by independent experts.

Since WordNet is semantically organized, the automatic text mining mechanisms could be of help to extract the desired information. These mechanisms mainly focus on the short-text similarities and conceptual relationships between terms and phrases in a text corpus.

A well-known approach for automatic extraction of information from text is first to design an algorithm that provides high-recall results. The next step would be a manual post-processing and revision by human experts that improves precision as well [23].

In this paper, we propose several automatic and semi-automatic methods for constructing ICT WordNet by means of the current data in Princeton WordNet. In these methods, we propose different criteria and algorithms to extract the ICT synsets from WordNet. In order to achieve this goal, we resort to various data such as *label*s in WordNet Domains, the definition of WordNet synsets (a.k.a *gloss*), existing semantic relations in Princeton WordNet, *LCS*, and S^3M similarity between WordNet definition and specialized words in Microsoft dictionary, and finally, *coordinate* synsets.

In order to evaluate the performance of our approaches, we use the precision measure as defined in Equation (1) and depicted in Figure 1. A query to

the WordNet retrieves a set of synsets, some of which are related to the query (e.g. are ICT synsets in our case) and others are irrelevant. Precision measures *purity* of the retrieved set and *recall* estimates its completeness.

$$precision = \frac{|related \cap retrievd|}{|retrieved|}$$
$$recall = \frac{|related \cap retrievd|}{|related|} \quad (1)$$

It is worth mentioning that an algorithm with a high precision returns significantly more relevant results than the irrelevant ones, while an algorithm with a high recall returns most of the relevant results along with a considerable amount of non-relevant ones. Therefore, we extract a set of high-recall results and then improve precision with the help of human expert revision.

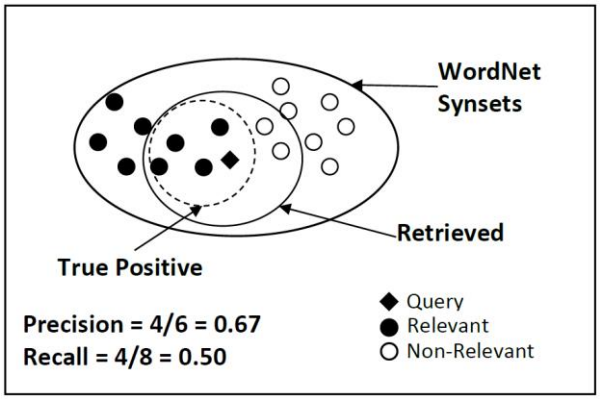

Figure 1. Precision and recall definition.

In the following sub-sections, we describe how each approach works, and present the results obtained.

3.1. Labels in Domain WordNet

Among the 167 current domain labels in WordNet Domains, we selected 12 ICT-related domains (Table 4). As a result, all the synsets belonging to these domains are labeled as the ICT synsets. Therefore, they can be directly added to our ICT WordNet.

3.2. ICT terms in synsets

In this method, we use a set of reference ICT words (shown in Table 4) to extract the appropriate synsets. In the first step, each synset including one of the words in Table 4 is considered as an ICT synset. In the second step, we also extract all synsets in semantic relations with synsets from the previous stage. All these synsets are candidates to be the ICT synsets. For the latter case, we use various semantic relations, as listed in Table 3.

Table 2. ICT domains in Domain WordNet.

ICT Domains		
Acoustics	Electronics	Graphic arts
Applied science	Electro technology	Telecommunication
Computer science	Engineering	Telegraphy
Electricity	Grammar	Telephony

Table 3. Semantic Relations in WordNet 3.0.

Semantic Relation	Example	
	To	From
hypernym	Home computer	Computer
hyponym	Input device	Keyboard
Instance Hypernym	Mozilla Firefox	Web browser
Instance Hyponym	Computer Scientist	John McCarthy
Part Holonym	Floppy Disk	Personal Computer
Part Meronym	Color TV Set	Color Tube
Member Holonym	Key	Keyboard
Member Meronym	Memory	Cell
Substance Holonym	Silicon	Transistor
Substance Meronym	Diode	Germanium
Entail	Send	Receive
Cause	Exception	Div by zero
Similar	Normal	Average
Also	Healthy	Faulty
Attribute	High-Speed	Computation
Domain Category	Screen Server	Computer Science
Domain Member Category	Computer Science	Screen Saver
Domain Region	Silicon Valley	California
Domain Member Region	California	Silicon Valley
Domain Usage	Windows	Trademark
Domain Member Region	Trademark	Windows

Table 4. Selected words in ICT domain.

ICT Words		
Audio	Magnetic	Protocol
Bit	Hardware	Signal
Byte	Interface	Technology
Buffer	Internet	Telephone
Cable	Microwave	Television
Compiler	Microprocessor	Telecommunication
Computer	Mouse	Software
Digital	Network	Video
Electrical	Operating System	Wire
Electronic	Printer	Wireless
File	Programming	Webpage
Folder		

Some well-known ICT words such as *bit*, *magnetic*, *mouse*, *file*, and *folder* severely decrease the precision while improving the recall. On the other hand, ignoring these words dramatically decreases the number of extracted ICT synsets. The reason is that these words have several meanings and usages in other domains. Therefore, we evaluated this method with the word list shown in Table 4 (first case) and another time with the same words without words *bit*, *magnetic*, *mouse*, *file*, and *folder* (second case).

3.3. ICT terms in gloss

Another approach is based upon the definition of the WordNet synsets denoted by gloss. In this method, after removing stop words, stemming, and removing repetitions, we use the percentage of the ICT terms (from Table 4) in the gloss as an indicator for ICT synsets, as defined in (2).

$$S = \frac{N_f}{N} \times 100 \quad (2)$$

Here, N is the number of total words in the gloss and N_f is the number of its ICT terms. A synset with $S > 20$ is considered as an ICT synset and is added to our ICT WordNet after confirmation by a human expert. Finally, synsets in semantic relation with the current (ICT) synset are also added to ICT WordNet. We report some results obtained by this method in Table 5.

3.4. ICT terms in neighbor synsets

The idea behind this method is that a synset is as an ICT synset if at least one of its neighboring synsets contains an ICT word. For a given synset, *S*, a neighbor is any other synset, *T*, which has some semantic relation to *S*.

In this method, we use the ICT terms from TechTerms Computer Dictionary [24]. The goal of TechTerms is to make computer terminology easy to understand. Some terms in TechTerms are commonly used and have definitions that are easy to understand. Others are less common and their definitions include a more advanced terminology. For this reason, each term includes a *tech factor*, ranging from 1 to 10. The terms with a low tech factor are basic and well-known terms, while the terms with high tech factors are more technical and are not used frequently. In our work, we use words with a tech factor from 1 to 5.

Generally, each set of words with a certain tech factor are divided into four categories. The first category includes words like *smartphone*, *drag and drop*, *double click*, and *gray scale*, which are not in WordNet. We delete all these words from our word set. Note that these words are assigned with a high tech factor. The second category contains words like *apple* and *Macintosh*, which exist in WordNet but their meaning is not related to the ICT domain. These words were also deleted from the word set. The third group contains common words with an ICT sense like *virus*, *memory*, and *character*. However, their usage in other domains is much more than in ICT. We removed the majority of these words as they resulted in extracting the non-ICT synsets. The fourth category including words like *computer*, *scanner*, and *Gigabyte* was the most appropriate one as its corresponding results were ICT with a high precision. Table 5 shows the results obtained by the *neighbor synsets* method. We categorize words into five groups according to their tech factor ranging from 1 to 5.

Table 5. Results for semantic relation / TechTerms method.

Tech Factor	Words	Synset	Precision
1	1148	712	0.69
2	1389	851	0.5
3	806	447	0.69
4	1023	616	0.80
5	930	537	0.85

The experiments show that the higher the tech factor, the lower the number of extracted ICT synsets. However, by increasing the tech factor, the precision also increases. The reason is that a word with a higher tech factor is more specific in the ICT domain and it unlikely has another sense in other domains. However, the results shown in Table 5 do not confirm this claim because the number of words with different tech factors is not the same. For example, the number of words with tech factor 2 is much more than the words with tech factor 1.

3.5. LCS-based approach

Longest common subsequence (*LCS*) is the problem of finding the longest subsequence common to two sequences. LCS between two sequences is computed using a dynamic programming procedure. Here, we extract the ICT information from WordNet using the Microsoft dictionary and LCS algorithm. The procedure is as follows: for each sentence S_d (definition of a word) in the Microsoft dictionary, we compute LCS between S_d and each synset definition throughout the WordNet. Then the criterion represented in (3) is taken to select the ICT synsets.

$$\mathrm{Sim}(\mathrm{d},\mathrm{w}) = \frac{\left|\mathrm{LCS}(\mathrm{S_w},\ \mathrm{S_d})\right|^2}{|\mathrm{S_w}| \times |\mathrm{S_d}|} \quad (3)$$

Here, S_w and S_d are the WordNet sentence and dictionary sentence, respectively. $|LCS(S_w, S_d)|$ represents the length of LCS between the WordNet synset gloss and the dictionary definition.

In order to have a unique and meaningful definition for each word, we performed some necessary pre-processing on the Microsoft dictionary. For example,

in the Microsoft dictionary, there were a large number of words with more than one definitions. There were also a lot of words for which the definition was only a reference to another entry in the dictionary. We also extracted the stems of all words in the synset definitions in WordNet as well as those in the Microsoft dictionary.

We evaluated the experiments based on the LCS similarity in different ways. In the experiment denoted by LCS-1, for each definition in the Microsoft dictionary, we extracted five most similar synset glosses according to the LCS measure. In another experiment denoted by LCS-2, we only rely on synset glosses, for which at least 10% of words are in the ICT domain. Then for each sentence in the Microsoft dictionary, we extract the five most similar synset glosses again according to the LCS measure. The results obtained are shown in Table 6.

This method takes all synsets into account including the ICT and non-ICT ones. Therefore, it extracts some ICT candidate synsets that are not extracted by the two previous methods. However, since the WordNet definition of a synset is very short compared to the long definitions in the Microsoft dictionary, the precision of the LCS method is not very promising.

3.6. S3M metric

The Sequence and Set Similarity Measure (S^3M) [25] is a similarity preserving function that captures both the order of occurrence of items in sequences and the constituent items of sequences. In other words, S^3M consists of two parts: one that quantifies the composition of the sequence (set similarity) and the other that quantifies the sequential nature (sequence similarity). Sequence similarity is defined as the order of occurrence of the item sets within two sequences. As in Equation (4), the length of the longest common subsequence (|LCS|) with respect to the length of the longest sequence determines the sequence similarity across two sequences.

$$SeqSim(A,B) = \frac{|\mathrm{LCS}(A,B)|}{\max(|A|,|B|)} \quad (4)$$

The set similarity, also known as the Jaccard similarity measure, is defined as the ratio of the number of common items and the number of unique item sets in two sequences. Therefore, the composition similarity of two series of A and B is measured as follows:

$$SetSim(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

Finally, S^3M is defined as follows:

$$S^3M(A,B) = p \times SeqSim(A,B) + q \times SetSim(A,B) \quad (6)$$

where, $p + q = 1$ and $p, q \geq 0$. Here, p and q determine the relative weights of sequence similarity and set similarity, respectively. Using different experiments, the weight parameters p and q *were tuned to 0.75 and 0.25, respectively.*

We use the S3M measure to extract the ICT-candidate synsets from WordNet.The procedure is as follows: for each word entry in the Microsoft dictionary, we compute the S3M measure between its definition and the gloss of each synset in WordNet. If similarity is higher than a pre-defined threshold, the synset is considered as ICT-candidate. The results obtained from the S3M criteria are shown in Table 6.

Note that, in some cases, several words appear in two sequences but not in the same order. The S^3M metric, specifically, helps to choose such synsets. In our experiment, we used 0.1 as the threshold measure for both the *LCS*- and S^3M-based methods.

3.7. Coordinate synsets

In this method, the coordinate relations between synsets are used for extracting the ICT information. As shown in Figure 2, the coordinate synsets are nouns or verbs that have the same hypernym.

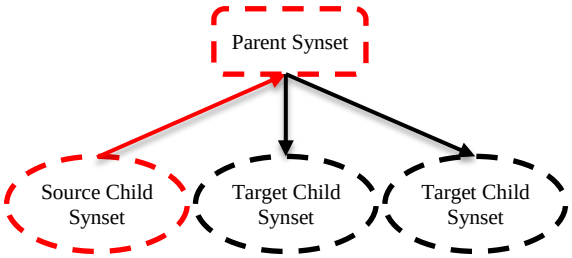

Figure 2. Co-ordinate synset lookup.

The coordinate-based method is only applied to the ICT synsets, which have been already extracted by one of the above-mentioned mechanisms. Our aim here is to verify whether the coordinate of each ICT synset is also ICT.

4. Experiments and results

Table 6 shows the overall results obtained from each one of the above mechanisms. We carried out two experiments, one time on the ICT synsets obtained by the mechanism proposed in Section 3.3 (denoted by

coordinate-exp1) and another time on the ICT synsets obtained from Section 3.4 (denoted by coordinate-exp2).

Table 6. Overall ICT data statistics.

No.	Method	Total Synsets	ICT Synsets	ICT Words	Precision
1	Synset ICT Words	2740	1726	2903	0.63
2	Gloss ICT Words	1662	1213	1789	0.73
3	ICT Ref-Words	3434	1923	3210	0.56
4	Neighbor Synsets	4455	3163	4148	0.71
5	LCS-1	13334	2000	2425	0.15
6	LCS-2	3450	1518	3160	0.44
7	S^3M	4590	1423	2247	0.31
8	Coordinate-Exp1	4475	1611	2588	0.36
9	Coordinate-Exp2	2485	1019	1702	0.41
	Total		**3625**	**4845**	

5. Analysis and discussion

We can observe in Table 6 that not only our mechanisms do not produce distinct results but also there is a high degree of overlapping among the results obtained by these approaches. We see that many of the extracted synsets and words are common. Totally, we extracted 3625 distinct synsets and 4845 distinct words using all our mechanisms.

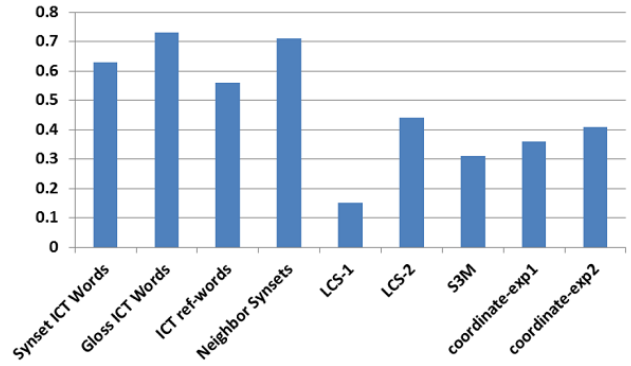

Figure 3. Precision obtained by each extraction mechanism.

As we can see in Table 6 and Figure 3, the first four approaches are more efficient in terms of precision. The precision obtained by the other mechanisms is not very promising as they generate a high false negative rate leading to a more non-ICT data.

We see in Figure 3 that the LCS-1 mechanism suffers from a very low precision. However, it extracts more ICT synsets (cf. Figure 4) compared to all the other mechanisms but one, which is a neighbor synsets approach.

As the result of all experiments, we conclude that an approach focusing on increasing precision extracts less ICT synsets. Therefore, for some approaches, we choose a more efficient reference set of words in order to increase the accuracy of results, while for some other mechanisms we use a threshold to achieve not only an acceptable precision but also extract a larger volume of ICT data.

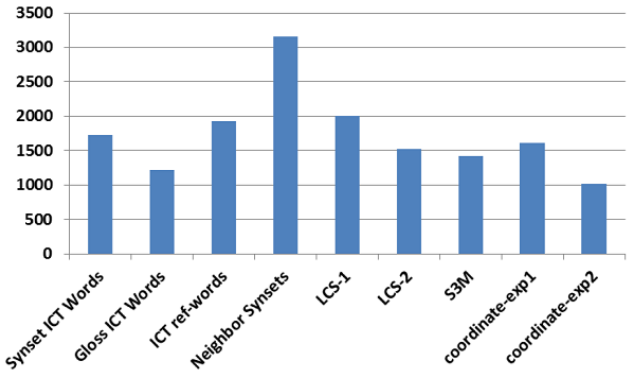

Figure 4. # of ICT synsets extracted by each mechanism.

6. Conclusion

In this paper, we have proposed several mechanisms to automatically extract the ICT data from the Princeton WordNet including synsets, words, and semantic relations between synsets.

In our first approach, each synset including at least one word of a reference ICT words was considered as an ICT synset. We also extracted all synsets in semantic relations with this synset. The accuracy of the results obtained by this method highly depends on the reference list of the ICT words. In our second approach, we extracted all synsets for which the ratio of ICT words in the gloss exceeded a pre-defined threshold. We observed that for this approach, an inverse relationship existed between the accuracy of results and the volume of the extracted ICT data. Therefore, we deployed a trial-and-error process to improve the results by selecting a proper reference list.

We also used the semantic relations to extract more ICT data. According to this mechanism, we used words in the neighboring synsets and ICT words from Tech Terms Computer Dictionary. By choosing words with higher *tech factor*, the accuracy of this mechanism was enhanced, while the number of extracted synsets was sharply decreased. This approach provided results with good precision. Using the *LCS* and S^3M criteria also helped expanding the

ICT data, even though they acquired a low precision. By these two similarity measures, it is easily possible to compute the similarity degree between the definition of an ICT word in a technical dictionary and the definition of any synset in WordNet. However, the precision of the results obtained was low due to the very short glosses in WordNet compared to the technical dictionaries. Finally, we proposed to use coordinate relation to explore more synsets in relation to an ICT synset. All the above-mentioned approaches help us to construct the first version of our ICT WordNet. We planned to create our own ICT corpus and to use various text mining algorithms in order to derive more ICT data to extend our ICT WordNet.

7. Acknowledgment

This work was partially supported by Iran Telecommunication Research Center (ITRC) under the contract No. 6614 as a part of Persian ICT WordNet project at Bu-Ali Sina University.

References

[1] Miller, G. (2011). Retrieved from Princeton WordNet, Available: http://wordnet.princeton.edu

[2] Miller, G. A. (1995). WordNet: a Lexical Database for English. Communications of the ACM, vol. 38, no. 11, pp. 39-41.

[3] Kremizis, A., Konstantinidi, I., Papadaki, M., Keramidas, G., & Grigoriadou, M. (2007). Greek WordNet Extension in the Domain of Psychology and Computer Science. Proceedings of the 8th Hellenic European Research Computer Mathematics and its Applications Conference (HERCMA), Economical University. Athens, Greece, 2007.

[4] Khazaei, A. & Ghasemzadeh, M. (2015). Comparing k-means clusters on parallel Persian-English corpus. Journal of AI and Data Mining. vol. 3, no. 2, pp. 203-208.

[5] Zahedi, M. & Arjomandzadeh, A. (2016). A new model for persian multi-part words edition based on statistical machine translation. Journal of AI and Data Mining. vol. 4, no. 1, pp. 27-34.

[6] Heshaam Faili, M. M. (2010). Automatic Persian WordNet Construction. 23rd International Conference on Computational Linguistics, Beijing, China, pp. 846-850.

[7] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S. & Assi, S. M. (2010). Semi Automatic Development of FarsNet; The Persian WordNet. Proceedings of 5th Global WordNet Conference. Mumbai, India, vol. 29.

[8] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic WordNet Project. Proceedings of the Third International WordNet Conference, Jeju Island, Korea, 2006, pp. 295-300.

[9] Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., & Oparin, I. (2002). Russnet: Building a Lexical Database for the Russian Language. Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation, pp. 60-64. Las Palmas, 2002.

[10] RussNet. (2012). Retrieved from RussNet Project, Available: http://project.phil.spbu.ru/RussNet/.

[11] Hitoshi ISAHARA, F. B. (2008). Development of the Japanese WordNet. Language Resources and Evaluation Conference. Retrieved from Japanese WordNet, Available: http://nlpwww.nict.go.jp/wn-ja/index.en.html, Marrakech, Morocco, 2008.

[12] Sagot, B. a. (2008). Building a Free French Wordnet from Multilingual Resources. OntoLex, 2008.

[13] Viberg, A., Lindmark, K., Lindvall, A., & Mellenius, I. (2002). The Swedish Wordnet Project. Proceedings of the Tenth EURALEX International Congress, Copenhagen, Denmark, 2002, pp. 407-412.

[14] Lam, K., Tarouti, F., & Kalita, J. (2014). Automatically Constructing Wordnet Synsets. 52nd Annual Meeting of the Association for Computational Linguistics. Baltiomore, USA, vol. 2, pp. 106-111.

[15] Yoon, A. e. (2009). Construction of Korean Wordnet Korlex 1.5. Journal of KIISE: Software and Applications, vol. 36, no. 1, pp. 92-108.

[16] Navigli, R. a. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. Artificial Intelligence, vol. 193, pp. 217-250.

[17] Nimb, S., & Pedersen, B. S. (2012). Towards a Richer Wordnet Representation of Properties. Language Resources and Evaluation Conference (LREC'12), 2012, pp. 3452-3456.

[18] Jurgens, D., & Pilehvar, M. (2015). Reserating the Awesometastic: An Automatic Extension of the WordNet Taxonomy for Novel Terms. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1459-1465.

[19] McCrae, J. P., Wood, I., & Hicks, A. (2017). The Colloquial WordNet: Extending Princeton WordNet with Neologisms. International Conference on Language, Data and Knowledge, Springer, Cham, 2017, pp. 194-202.

[20] Rusert, J. (2017). Language Evolves, So Should WordNet: Automatically Extending WordNet with the

Senses of Out of Vocabulary Lemmas. PhD thesis: University of Minnesota.

[21] Magnini, B., & Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. Language Resources and Evaluation Conference (LREC'00), pp. 1413-1418.

[22] Poprat, M. E. (2008). Building a BioWordNet by using WordNet's Data Formats and WordNet's Software Infrastructure: a Failure Story. ACM Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp 31-39.

[23] Aggarwal, C. C., & Zhai, C. X. (2012). Mining Text Data. Springer Science & Business Media, 2012.

[24] TechTerms. (2012). TechTerms. Retrieved from TechTerms: http://www.techterms.com/

[25] Kumar, P., Raju, B. S., & Krishna, R. P. (2011). A New Similarity Metric for Sequential Data. Exploring Advances in Interdisciplinary Data Mining and Analytics: New Trends: New Trends. IGI Global, pp. 233.

نشریه هوش مصنوعی و داده کاوی

ساخت خودکار وردنت فارسی در حوزه‌ی فاوا به کمک وردنت پرینستون

اکرم احمدی طامه، محمد نصیری* و محرم منصوری زاده

گروه کامپیوتر، دانشکده مهندسی، دانشگاه بوعلی سینا، همدان، ایران.

ارسال ۲۰۱۶/۱۱/۰۸؛ بازنگری ۲۰۱۷/۰۸/۱۴؛ پذیرش ۲۰۱۸/۰۴/۰۵

چکیده:

وردنت یک پایگاه داده‌ی لغوی بزرگ شامل اسم، فعل، صفت و قید در زبان انگلیسی است که واژگان هم‌معنی را در مجموعه‌هایی به نام ترادف دسته-بندی می‌کند. هر ترادف بیانگر یک مفهوم جداگانه است که توسط روابط ساختاری و معنایی با ترادف‌های دیگر مرتبط می‌شود. از وردنت در ابهام زدایی واژگان، بازیابی اطلاعات و ترجمه‌ی متون استفاده می‌شود. در این مقاله چندین روش خودکار برای استخراج دادگان حوزه‌ی فناوری اطلاعات و ارتباطات (فاوا) از وردنت پرینستون ارائه می‌شود. این اطلاعات به دادگان وردنت فارسی ما اضافه می‌شوند. مزیت روش‌های خودکار، کاهش دخالت عامل انسانی و در نتیجه سرعت بخشی به فرایند توسعه وردنت دوزبانه فاوا است. در اولین روش پیشنهادی، با جستجوی مجموعه محدودی از واژگان فاوا در متن تعریف یک ترادف، تعلق آن به حوزه فاوا بررسی می‌شود. در روش دوم، ترادف‌هایی که در رابطه معنایی با ترادف‌های فاوا باشند، استخراج می‌شوند. روش بعدی از دو معیار شباهت LCS و S^3M برای سنجش میزان شباهت تعریف یک ترادف با تعریف واژگان در فرهنگ واژگانی مایکروسافت بهره می‌گیرد. آخرین روش، ترادف‌های برادر در وردنت را بررسی می‌کند. نتایج بدست آمده نشان می‌دهد که روش‌های پیشنهادی قادرند دادگان فاوا را با دقت قابل قبولی از وردنت پرینستون استخراج نمایند.

کلمات کلیدی: وردنت، فاوا، ساخت خودکار، ترادف، رابطه معنایی.