Natural Language Engineering **23** (1): 53–91. © Cambridge University Press 2015
doi:10.1017/S1351324915000376

Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features

ABDULGABBAR SAIF, MOHD JUZAIDDIN AB AZIZ and NAZLIA OMAR

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor, Malaysia
e-mail: agmss79@siswa.ukm.edu.my, juzaiddin@ukm.edu.my, nazlia@ukm.edu.my

(*Received 3 March 2014; revised 16 September 2015; accepted 18 September 2015; first published online 21 October 2015*)

Abstract

The alignment of WordNet and Wikipedia has received wide attention from researchers of computational linguistics, who are building a new lexical knowledge source or enriching the semantic information of WordNet entities. The main challenge of this alignment is how to handle the synonymy and ambiguity issues in the contents of two units from different sources. Therefore, this paper introduces mapping method that links an Arabic WordNet synset to its corresponding article in Wikipedia. This method uses monolingual and bilingual features to overcome the lack of semantic information in Arabic WordNet. For evaluating this method, an Arabic mapping data set, which contains 1,291 synset–article pairs, is compiled. The experimental analysis shows that the proposed method achieves promising results and outperforms the state-of-the-art methods that depend only on monolingual features. The mapped method has also been used to increase the coverage of Arabic WordNet by inserting new synsets from Wikipedia.

1 Introduction

Lexical knowledge sources are considered as the cornerstone of many natural language processing tasks, such as measuring semantic similarity, named entity (NE) recognition, and word-sense disambiguation. The performance of knowledge-based tasks depends mainly on the quality and quantity of the exploited knowledge sources (Zhang, Gentile and Ciravegna 2012). Accordingly, the need for building a highly efficient semantic lexicon is becoming increasingly important. For the English language, there are several electronically available knowledge sources, such as the Longman Dictionary (Paul 1978), Roget's thesaurus (Roget 1911), and WordNet (Fellbaum 1998). The Princeton WordNet (PWN) is a large lexical database for English consisting of sets of synonyms (synsets) and connections (lexical relations between the synsets). Due to its accessibility and quality (manual annotations by lexicographers), PWN is the most extensively used knowledge source in information retrieval and computational linguistics. Moreover, the structure of PWN has been used to build WordNets for other languages. However, the manual building of a lexical database is a laborious and time-consuming task because it requires

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

lexicographic expertise to analyze the variations in using words syntactically and semantically in their natural language contexts.

Wikipedia has been exploited as a semi-structured knowledge source since it contains a huge number of concepts (Wikipedia articles), categories, monolingual semantic relations (links between articles and links between an article and its categories), and cross-lingual relations (links among the articles in different languages). Promising results have been achieved for many natural language processing tasks, such as measuring semantic relatedness (Gabrilovich and Markovitch 2007; Milne and Witten 2008; Hassan and Mihalcea 2009; Hassan and Mihalcea 2011), word-sense disambiguation (Mihalcea 2007), and NE recognition (Cucerzan 2007). As a result of the continuous contributions in updating Wikipedia by many volunteers, it has become a high coverage knowledge source that covers multiword expressions and NEs: people, locations, organizations, and so on.

Despite the differences between Wikipedia and WordNet in terms of the constructors, purpose of construction, and the degree of structuring and coverage of semantic contents, the basic unit of Wikipedia (an article) can appear like a synset in WordNet. In other words, for some features in a Wikipedia article there are similar features in WordNet. For instance, the title, redirects, first paragraph, and categories of an article resemble the sense, synonyms, gloss, and taxonomical relations of a synset in WordNet. Recently, researchers have incorporated the features of Wikipedia and the features of WordNet to build new knowledge source that has more semantic efficiency than the existing sources. Moreover, the coverage of WordNet can be enriched by adding concepts from Wikipedia. The mapping establishes the links between WordNet synsets and Wikipedia articles on the basis of the similarity of the features in the two sources. However, the major challenge in this task is how to address the ambiguity issue (polysemy and homonym) for the lemmas (single word or multiword) in the identifiers of WordNet synsets (sense and synonyms) and Wikipedia articles (title and redirects).

Several studies (Niemann and Gurevych 2011; Navigli and Ponzetto 2012; Matuschek and Gurevych 2013; Pilehvar and Navigli 2014) on English sources have considered the ambiguity issue using the semantic similarity of monolingual features. In those studies, the similarity between a synset and an article is computed using two main techniques: overlap-based and WordNet-based techniques. The overlap-based technique depends on the number of words that are shared in the text representations of a synset and an article. The WordNet-based technique has been introduced to overcome the limitations of the overlap-based technique (which considers words as if they are independent elements) by using WordNet as an ontology to model the semantic similarity based on the semantic relations among synsets. According to Zesch and Gurevych (2010), the degree of the completeness of the semantic source is important for the performance of the semantic measures. Due to the high degree of the completeness of PWN compared to other WordNets, the WordNet-based technique achieved encouraging results both in text similarity and in the mapping of English. However, non-English WordNets with poor semantics cannot provide enough semantic evidence for addressing the ambiguity and synonymy issues in mapping.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

In this paper, to cover the gap in the WordNets with poorer semantics, we investigate two important issues that have not been addressed in the previous mapping works. First, the existing works depend only on monolingual features to semantically link the concepts in the different sources. The feasibility of using cross-lingual features has not been investigated for the mapping task. Second, the semantic similarity in the previous works is typically measured based on the semantics in the WordNet (a single knowledge source). Although the structural elements in Wikipedia are used to represent each article as a set of words, the Wikipedia-based semantics have not been used to model the semantic relatedness between the synset and the article from the evidence that can be extracted from Wikipedia. Therefore, we propose the synset–article mapping method, which links a synset unit in Arabic WordNet (AWN) to its corresponding concept in Arabic Wikipedia (AWP), depending on the monolingual and bilingual (Arabic–English) features. In this method, AWN and AWP are exploited together (rather than WordNet only as in the previous works) as the semantic proxies of the semantic similarity measures. Based on the semantic proxies of AWN and AWP, two semantic measures (synset-based and article-based) are introduced to model the similarity between the monolingual features of the synset and article. Furthermore, the proposed method is adapted to extend the coverage of AWN from the concepts in Wikipedia.

The motivation behind incorporating bilingual features with monolingual features in the mapping task can be summarized as follows. From the perspective of the disambiguation, the senses of some ambiguous words in one language have distinct meanings in another language. Thus, far less data in bilingual features can be useful for determining the correct sense of an ambiguous word in a language for cheaper cost. For instance, the bilingual representation [*century*, قرن provides complementary evidence for the semantics of the ambiguous word 'قرن', which has two senses (*100 years* and *one of the bony outgrowths on the heads of certain ungulates*) in AWN. The bilingual representation also can easily resolve the ambiguity of some words in the two languages. For example, the English word 'back' with part of speech *noun* has six senses in PWN, and the Arabic word 'ظهر' has three meanings (*part of the body*, *noon*, and *pray*), but the pair of words [back, ظهر] carries only one semantic meaning in both languages. On the other hand, the bilingual features can play a distinct role for handling the synonymy issue in matching concepts from different knowledge sources. In the Arabic language, there is an incompatibility in the terms for some concepts across sources such as WordNet and Wikipedia. This incompatibility is due to the incompleteness of the synonyms in AWN, the multiplicity of dialects, and Arabization problems. Arabization is the linguistic transfer of foreign terms that refer to concepts (objects or events) from a language to Arabic. Unlike translation, which assumes that a term in one language has an equivalent term in the other language, Arabization (Al-Asal and Smadi 2012) focuses on tackling the terms in the other language that have no direct equivalent in Arabic (the untranslatability problem), using derivation, compounding, word borrowing, and blending. Due to the multiplicity of coiners for Arabic across Arab countries, most untranslatable terms have been transferred in several forms to Arabic. For example, 'biome' and 'pixel' have been transferred to Arabic in different lemmas as ('بيوم' versus 'مجال حيوي بيئي')

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

and (‘نقطة ضوئية’ versus ‘عنصورة’) in AWN and AWP. The bilingual identifiers of these terms are similar in both English WordNet and Wikipedia. Based on cross-lingual features, we hypothesize that exploiting bilingual features with monolingual features can yield important improvements for the ambiguity issue as well as the mapping task.

The remainder of the paper is organized as follows. Section 2 presents the previous related works in mapping WordNet entities to Wikipedia articles. Section 3 explores the knowledge sources (AWN and Wikipedia) used in the mapping method. Section 4 provides the Wikipedia-based method for Arabic semantic representation. Section 5 introduces the proposed method for mapping AWN synsets to Wikipedia concepts. Section 6 reports the mapping results for all nominal synsets in AWN, while Section 7 shows the evaluation of the results in several experiments for the proposed method. An extension of the method to add new synsets to the AWN database is introduced in Section 8. Eventually, Section 9 provides the conclusions and suggestions for future work.

2 Previous related works

Automatic alignment of WordNet and Wikipedia has been proposed in recent years to enrich the semantics in WordNet or to build a new knowledge source. Based on the types of entities in Wikipedia, this alignment can be classified into two kinds: category–synset and article-synset. For the category-synset alignment, Suchanek, Kasneci and Weikum (2007) introduced a linkage algorithm to link a Wikipedia category to a taxonomy WordNet synset for building a new knowledge-based ontology known as *Yet Another Great Ontology* (YAGO). For the given category, its headword was determined using shallow noun-phrase parsing and the WordNet synsets that match this head word. These were extracted as the candidate set for a given category. If a candidate set contained only one synset, it was selected as the class for the given category without any further processing. Otherwise, the simplest technique, word-sense disambiguation, was applied to select the most frequent sense of the headword in WordNet as the class for the given category. For the word-sense disambiguation, Toral, Munoz and Monachini (2008) used the overlap measure between the titles of the articles belonging to the given category and the instances of each synset in the candidate set.

The main challenge in article–synset alignment is how to address the ambiguity of the candidates when selecting a suitable article for the given synset. This issue is a special type of word-sense disambiguation. Several studies have considered how to select the most relevant article for a synset by measuring the similarity. For instance, Ruiz-Casado, Alfonseca and Castells (2005) proposed an algorithm to assign an article automatically in sample English Wikipedia to the corresponding synset in WordNet. In this algorithm, the article and the synset were represented as vectors using the content of the article and the gloss with the hierarchical relations of the synset. Then, the similarity measure between the article and the synset was computed using the dot product and cosine function between the two vectors of the article

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

and the synset. Finally, the synset with the highest similarity was selected to assign the given article to this synset.

Ponzetto and Navigli (2010) proposed a mapping method that links a Wikipedia article to a WordNet sense. In this method, the Wikipedia article was represented as a disambiguation context containing the set of words extracted from the following features: title, outgoing links, and categories. On the other hand, the disambiguation context of the WordNet sense was extracted as the set of words from the following: synonyms, hypernym/hyponym, sisterhood, and gloss. The gloss-based technique was used to measure the similarity between the sense and article as the traditional text overlap (intersection) score between their disambiguation sets. For a given article w, the mapping algorithm was used to select one sense s from the set of senses of the lemma w in WordNet depending on a maximization of the conditional probability $p(s|w)$ over the senses. To improve the overlap measure, Navigli and Ponzetto (2012) introduced a graph-based technique to compute this score depending on the WordNet semantic graph, which was constructed from the lexical relations among the synsets. For a given article w, its disambiguation context was constructed from the semantic features in Wikipedia as the bag-of-words set. Then, the senses of each word in the disambiguation context of the article were retrieved from WordNet to build the disambiguation graph with these senses as the nodes. For each node, the depth-first search algorithm was used for the WordNet graph to discover the paths (if any) of the given maximal length from this node to other nodes in the disambiguation graph. The edges and intermediate nodes found in each path discovered were added to the disambiguation graph. Finally, the semantic similarity between the article and the sense was measured based on the lengths of paths between the synset*s* and each node in the subgraph. The evaluation results showed that the graph-based technique achieves better accuracy (82.7 per cent) than the gloss-based technique (80.9 per cent).

Wolf and Gurevych (2010) proposed a method for building a wide sense inventory from the knowledge included in Wikipedia and WordNet. In this method, the structure of Wikipedia titles was exploited to retrieve article candidates from Wikipedia for a WordNet synset. For a given term t, the Wikipedia article was extracted as the candidate according to these constraints: the title of the article matches the term, the article title has the form t_(*description tag*), the term t redirects to the article, or the form t_(*description tag*) redirects to the article.

Niemann and Gurevych (2011) proposed a multiple-based method to align none, one, or more than one alignment of a Wikipedia article for a given WordNet synset. This method has two main steps: candidate extraction and candidate alignment. In the candidate extraction step, the Wikipedia articles were retrieved as the candidate set according to the title constraints (Wolf and Gurevych 2010) for each term in the synset. For the candidate alignment step, the synset and its article candidates were represented as the bag-of-words extracted from the semantic information in WordNet and Wikipedia, respectively. Then, a personalized version of the PageRank algorithm (Agirre and Soroa 2009) was used to measure the semantic similarity between the WordNet synset and each article in the candidate set. The n-fold cross-validation technique was applied on the training set to estimate the threshold that maximizes

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

performance. Finally, the semantic similarity with the threshold was used to classify the synset-article pair as an alignment or a non-alignment. This method achieved an encouraging performance of 94.5 per cent accuracy on the well-balanced dataset.

Fernando and Stevenson (2012) used the information retrieval technique for extracting article candidates for a given synset based on entire articles in Wikipedia. The purpose was to extend the candidate set retrieved using the title constraints (Wolf and Gurevych 2010). For mapping a WordNet synset to its corresponding articles, two similarity measures were computed between the synset and each article in the candidate set. The first measure was the normalized intersection-based measure between the text representation set of the article and the representation set of the synset. The second measure was the matching-based score between the expressions in the synset and the title of the article.

Recently, a graph-based technique (Matuschek and Gurevych 2013; Pilehvar and Navigli 2014) was introduced to align between two semantic graphs. Based on the semantic relations in each knowledge source, a semantic graph is constructed, in which the nodes represent the concepts and the links between the nodes are the semantic relatedness between these concepts. For alignment between WordNet and Wikipedia, the semantic relatedness was determined based on the direct WordNet relations and the category links with the hyperlinks in the first paragraph of each article. For two given separate semantic graphs, Matuschek and Gurevych (2013) proposed the Dijkstra–WSA method to align between the nodes in these graphs by discovering the bridges between the graphs (in the initialization of the mapping). In this method, the nodes of the monosemous lemmas in the two sources are used to establish the link 'bridges' between the two resource graphs, in which an edge is added from a node in the first graph to its corresponding node in the second graph. After initialization, the mapping algorithm iterates over all the nodes of the ambiguous lemmas in the first graph that have not been linked to the corresponding nodes in the second graph. For each node s_i in the first graph, its candidates are extracted based on lemma matching to the set of nodes from the second graph. To select the correct concept from the candidates, the Dijkstra algorithm is used to determine the candidate t_i with the shortest path as the corresponding concept for this node s_i. Then, the new bridge (edge) is created from s_i to t_i, which captures additional semantic evidence about the relations of the remaining nodes.

Pilehvar and Navigli (2014) introduced the structural similarity to measure the semantic relatedness between two nodes from different graphs based on the structure of the semantic graphs of the resources. For each node in a graph, the Personalized PageRank (PPR) algorithm extracts the structural signature of the node as a representation vector. The structural similarity between two nodes from the two graphs depends on determining a common ground between the two representation vectors, which have different dimensions. The concepts of the monosemous lemmas in the two sources are used as a common ground to unify the representation vectors of the nodes in the two graphs, in which all the vectors have the same dimension (the number of these concepts). Thus, the structural similarity between two nodes is measured according to the similarity between their unified vectors. The experimental

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

results showed that this work (95.9 per cent accuracy) outperformed the Dijkstra–WSA method (93 per cent accuracy) on the same data set.

The bilingual features from different resources have been exploited in semi-automatic methods (Atserias *et al.* 1997; Vossen *et al.* 2012; Pradet, de Chalendar and Desormeaux 2014) for building and extending WordNets for languages other than English. There have only been a small number of previous studies aimed at enriching AWN entities from available resources automatically. Alkhalifa and Rodríguez (2008) introduced a method for NE extraction to enrich the NE coverage of AWN using four knowledge sources: PWN, English Wikipedia, AWP, and AWN. The first step in this method is to extract the NEs (candidate instances) from PWN that have not Arabic counterparts in the AWN. Then, the mapping technique was applied to select the corresponding NEs of the candidates from English Wikipedia. The title of the corresponding concept of English NE in Wikipedia was returned as the Arabic Named Entity. The experimental results showed that the accuracy of this method was 93.3 per cent. Abouenour, Bouzoubaa and Rosso (2010) used the YAGO ontology to enrich the NEs in AWN. The entities in YAGO that were in English were translated into Arabic using the *Google Translation API*. The translated entities were transformed into AWN synsets by linking the relation facts in YAGO to PWN's synsets, which in turn were connected with Arabic synsets, or by mapping the relation facts in YAGO directly to AWN synsets. Then, the automatic validation of NE links to corresponding AWN synsets was used to eliminate incorrect mappings as well as wrongly translated entities. The experimental results on the three million YAGO entities showed that it was possible to keep 4,33,339 instances that were connected with 2,366 corresponding AWN synsets. However, such previous studies focused only on enriching NEs in AWN using English features only.

The concepts in each language can be classified based on their semantics into two types: universal (the concepts that are fully equivalent and interchangeable across languages) and specific to a language (culture-related concepts). According to Kashgary (2011), there are several Arabic culture-related words, terms and concepts (for Islam, Arabic customs, and social life) that have no direct English equivalents or have not been lexicalized in English yet. Due to the differences between Arabic and English at many linguistic levels, a mapping using only English features will retain the semantics' characteristics of the English language and the specificity of Arabic will be missed.

3 Knowledge sources

3.1 Arabic WordNet

AWN (Elkateb *et al.* 2006; Rodríguez *et al.* 2008) is a freely available lexical database for standard Arabic that has been constructed with the same methodology as EuroWordNet (Vossen 1998). AWN is one of the most promising resources. It is structured as a lexical semantic ontology and it can be used for several Arabic language processing applications. Like other WordNets, AWN was designed to represent the core unit in WordNet called the synset. A synset unit is a set of synonyms that are grouped together to describe the meaning (sense) of a word for

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 1. *Statistics for Arabic WordNet*

POS	Synsets	Words
Noun	7,960	15,890
Verb	2,538	6,083
Adjective	661	1,238
Adverb	110	262
Total	11,269	23,481

a part of speech. Synsets are divided based on the parts of speech into four types: noun, verb, adjective, and adverb. Table 1 shows the number of synsets and words in AWN with their part of speech types. In this work, we focus only on the nominal synsets (with part of speech *noun*) because Wikipedia concepts are nouns. Arabic nominal synsets have been classified manually by lexicographers into NE synsets (1,147) and linguistic-term synsets (6,813).

The synsets in AWN are linked to each other through the different types of lexical relations that are used to represent knowledge as a semantic ontology. AWN contains 18,340 links, distributed across17 types of lexical relation such as *hyponym*, *antonym*, *holonym*, and *related to*. The hypernym (*IS-A*) is a hierarchical relation that is used to build the semantic taxonomy of the noun and verb synsets. In the semantic taxonomy, the concepts are organized according to the specificity notion (from general concept to specific concepts) as a directed acyclic graph for the noun or verb synsets. In path-based and information content methods for measuring the semantic similarity, the measures are computed between two nodes (synsets) based on the positions of these nodes in this semantic taxonomy. In this paper, the term *non-hierarchical* (NH) relations refers to all lexical relations (except *IS-A*) in the AWN database. Due to AWN being constructed in parallel with PWN, there is a bilingual semantic relation *equivalence* between an Arabic item and a PWN item, in which each Arabic synset is connected to a corresponding synset in English.

The gloss is a very important part of WordNet, containing a short textual definition of a synset. Unlike English WordNet, of 7,960 Arabic nominal synsets in AWN, only 189 (2.3 per cent) synsets have textual definitions (direct glosses). To overcome this limitation, for the other synsets (97.7 per cent), the glosses are extracted from the semantic taxonomy of AWN on the basis of the hierarchical relations among the synsets. For each node (synset), the gloss is generated by concatenating the lemmas in its related concepts (ancestors, children, and siblings) as the text (sequence of words). The ancestors of the synset are extracted as the nodes in the paths from the node that corresponds to this synset to the *root* of the taxonomy. The nodes at different levels in the semantic taxonomy have different numbers of possible ancestors, as anode in the lowest level has more ancestors than anode in the top level. Since anode can have more than one path to the root (multiple hypernym relations), the ancestors are extracted from each path. For example, the ancestors of the synset 'أَتْرُوبِين' 'atropine$_1^n$' are (سَمّ,تِرْيَاق,مَادَّة,دَوَاء,مُخَدِّر,شِبْهُ قَلْوِيّ,مُوَسِّعٌ للحَدَقَة,مُرَكَّب,مُضادُّ التَّشَنُّج,كَيْنُونَة,مُسَبِّب,مُرَكَّب عُضْوِيّ,عَامِل,أَتْرُوبِين,دَوَاء) (*poison*$_1^n$, *antidote*$_1^n$,*alkaloid*$_1^n$, *mydriatic*$_1^n$,*antispasmodic*$_1^n$,

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

remedy$^{n}_{2}$,*medicine*$^{n}_{2}$,*drug*$^{n}_{1}$, *organic compound* $^{n}_{1}$,*compound* $^{n}_{1}$, *agent* $^{n}_{2}$, *causal agent* $^{n}_{1}$, *substance* $^{n}_{1}$, *entity* $^{n}_{1}$), which are extracted from the following paths from the node that corresponds to the synsets for the root in the AWN taxonomy:

P1
أَتْرُوبِين←سَمّ←مَادَّة←كَيْنُونَة←Root
atropine$^{n}_{1}$ →poison$^{n}_{1}$→substance$^{n}_{1}$→entity$^{n}_{1}$→Root

P2
أَتْرُوبِين←تِرْيَاق←دَوَاء←مُخَدِّر←عَامِل←مُسَبِّب←كَيْنُونَة←Root
atropine$^{n}_{1}$→antidote$^{n}_{1}$→remedy$^{n}_{2}$→medicine$^{n}_{2}$→drug$^{n}_{1}$→agent$^{n}_{2}$→causal agent$^{n}_{1}$→ entity$^{n}_{1}$→Root

P3
أَتْرُوبِين←شِبْهُ قِلْوِيّ←مُرَكَّب عُضْوِيّ←مُرَكَّب←مَادَّة←كَيْنُونَة←Root
atropine$^{n}_{1}$→alkaloid$^{n}_{1}$→organic compound$^{n}_{1}$→compound$^{n}_{1}$→substance$^{n}_{1}$→ entity$^{n}_{1}$→Root

P4
أَتْرُوبِين←مُوَسِّعٌ للحَدَقَة←مُخَدِّر←عَامِل←مُسَبِّب←كَيْنُونَة←Root
atropine$^{n}_{1}$ →mydriatic$^{n}_{1}$→drug$^{n}_{1}$→agent$^{n}_{2}$→causal agent$^{n}_{1}$→ entity$^{n}_{1}$→Root

P5
أَتْرُوبِين←مُضادُّ التَّشَنُّج←دَوَاء←مُخَدِّر←عَامِل←مُسَبِّب←كَيْنُونَة←Root
atropine$^{n}_{1}$→antispasmodic$^{n}_{1}$→medicine$^{n}_{2}$→drug$^{n}_{1}$→agent$^{n}_{2}$→causal agent$^{n}_{1}$→ entity$^{n}_{1}$→Root

The semantic associations between a node and its ancestors vary according to the position of the node in the taxonomy, in which a deeper node is more semantically related to its nearest ancestors than a shallower node. For instance, the direct hypernym of a shallower node (*depth* = *2*) that corresponds to the synset 'هَالُوجِين' 'halogen$^{n}_{1}$' is the synset 'جَمَاعَة' 'group$^{n}_{1}$', which contains the general words 'زمرة,تجميع,مجموعة'. In the first path of the node 'أَتْرُوبِين', which contains the ancestors (كَيْنُونَة,مَادَّة,سَمّ), the lemmas in the nearest nodes 'مَادَّة' and 'كَيْنُونَة' are considered general because they can be extracted for several synsets (e.g. the synset 'شَرَاب' 'beverage' has the path "شَرَاب←طَعَام←مَادَّة←كَيْنُونَة"). On the other hand, the path from the node that corresponds to the synset 'ثَوْر' 'ox$^{n}_{1}$' to the root includes 15 ancestors as follows:

ثَور←بَقَر←بَقَرِيّ← بَقَرِيّات←مُجْتَرّ←ذُو حَوَافِر مُزْدَوِج الأَصَابِع←ذُو حَوَافِر← مَثِيمِيَّات← الثَّدْيِي← فَقَّارِي← حَبْلِي← حَيَوَان← المُتَعَضِّي←كَائِن حَيّ←جِسْم←كَيْنُونَة←Root

ox$^{n}_{1}$→cattle$^{n}_{1}$→bovine$^{n}_{1}$→bovid$^{n}_{1}$→ruminant$^{n}_{1}$→even-toed ungulate$^{n}_{1}$→ ungulate$^{n}_{1}$→ placental$^{n}_{1}$→ mammal$^{n}_{1}$→ vertebrate$^{n}_{1}$→chordate$^{n}_{1}$→animal$^{n}_{1}$→ organism$^{n}_{1}$→living thing$^{n}_{1}$→object$^{n}_{1}$→entity$^{n}_{1}$→Root

The nearest ancestors of the synset are the nodes 'بَقَر', 'بَقَرِيّ', 'بَقَرِيّات', and 'مُجْتَرّ', which contain the lemmas related to this synset. Therefore, the number of ancestors of anode is determined based on the depth of this node in the semantic hierarchy. For a node v with its paths to the root, the number of nearest ancestors from each path p that should be included in the gloss is defined by the following formula:

$$N_{\text{ancestors}}(p_v) = \text{Round}\left(\frac{\text{Length}(p_v) \times 4}{\text{Max}_{\text{depth}}}\right) \quad (1)$$

where $\text{Max}_{\text{depth}}$ is the depth of the AWN taxonomy,[1] $\text{Length}(p_v)$ returns the length of the path p_v in terms of its number of edges, and $\text{Round}(x)$ is a function that rounds a number x to the nearest integer (e.g. $\text{Round}(1.5) = 2$, $\text{Round}(1.4) = 1$). We used that number in the numerator in order to make the maximum of number

[1] The maximum depth of an AWN taxonomy is 16.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

of selected ancestors equal to four synsets which have strong semantic relationships with the specific node. For each path from each node to the root, this formula returns a number ranging between 0 and 4. The numbers of nearest ancestors of the paths of the synset 'اتْرُوبِين' are 1, 2, 2, 2, and 2, respectively. Thus, the nearest ancestors of this synset are 'سَمّ' *poison*$^{n}_{1}$, 'شِبْهُ قِلْوِيّ' *alkaloid*$^{n}_{1}$, 'مُرَكَّب عُضْوِيّ' *organic compound*$^{n}_{1}$, 'تِرْيَاق' *antidote*$^{n}_{1}$, 'دَوَاء' *remedy*$^{n}_{2}$, 'مُوَسِّعٌ للحَدَقَة' *mydriatic*$^{n}_{1}$, 'مُخَدِّر' *drug*$^{n}_{1}$, and 'مُضادُ التَّشَنُّج' *antispasmodic*$^{n}_{1}$.

The children of a synset are the nodes (synsets) that are directly subsumed (hyponymy relations) by this synset in the taxonomy. Since most nodes[2] are leaves in the AWN taxonomy (they do not have children), the siblings of each node are extracted as the related concepts instead of the children. The siblings of a synset are the nodes with the same parent in the synset. There can be different numbers of children or siblings across the nodes in the taxonomy. Thus, each node is represented by at most four nodes from its children or siblings. For a synset that has more than four related nodes, based on the specificity of the nodes in the taxonomy, we select the top four nodes from its children or siblings to be the related concepts of this synset. The specificity of a node v is defined based on the length of the shortest path between the node and its leaves by the following formula:

$$\text{specificity}\,(v) = \frac{1}{\min\limits_{c \in leaves(v)} \{\text{length}\,(v, c)\}} \tag{2}$$

where the function *length* returns the number of nodes in the shortest path between two synsets in the taxonomy.

3.2 Arabic Wikipedia

Wikipedia is a goldmine containing a massive amount of structured and unstructured information that is increasing and being pruned continuously (Medelyan *et al.* 2009). As a knowledge resource, the basic unit in Wikipedia is the article, which describes the single concept named by the unique title. The organization of Wikipedia addresses two problems due to the uniqueness of the title: polysemy and synonymy. Polysemy occurs when a word or phrase has more than one sense (meaning). For example, the word *bar* is ambiguous because it refers to several concepts: *bar (law)*, *bar (music)*, *bar (counter)*, *bar (establishment)*, and *bar (landform)*. Synonymy is when different words or phrases refer to the same concept. For this problem, the redirect links and anchor hyperlinks can be considered to be the inventory of the synonyms and variation forms of the specific concept (Niemann and Gurevych 2011). For example, the Wikipedia concept *Psychology* is redirected from the following terms: *hyde event*, *Human trait*, *Psychological*, *Pyhscology*, *Phsycology*, and *Psychological terms*.

Wikipedia concepts are categorized in a hierarchical organization ranking from highly specific to highly generic classes. Each Wikipedia concept belongs to one or more categories. These categories can be represented as the semantic relations that describe the meaning of the Wikipedia concept in terms of the classifications.

[2] Among 7960 nominal synsets, the number of leaves is 5217.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 2. *Statistics for Arabic Wikipedia*

Articles	Statistics	Categories	Statistics
Non-useful pages	7,587	Administrative categories	12,359
Concepts	1,57,173	Categories	38,018
Redirects	1,50,370	Links between categories	95,175
Links between articles	5,662,744	Links between categories and articles	6,65,953

Table 3. *Examples of NE candidates for linguistic-term synsets*

Arabic words	# of senses	Article candidates	Translation	Type of NEs
عين	4	عين_(بعلبك)	Ain (Beqaa)	Location
		عين_(النبطية)	Ain(Nabatieh)	Location
		عين_(نادي)	Ain(Football Club)	Organization
		عين_(مدينة)	Ain(city)	Location
		عين_(صيدا)	Ain(Sidon)	Location
دنيا	2	دنيا_(مسلسل)	Dunia(Serial)	Media
		دنيا_(فيلم)	Dunia(film)	Media
		دنيا_(ممثلة)	Dunia(actress)	Person

For example, the concept *Chemistry* belongs to three categories *Chemistry*, *Natural sciences*, and *Physical sciences*. Wikipedia has categories that are used for maintenance due to the crowdsourcing process of creating and editing articles. For example, the *category: page need revise*, *Articles with dead external links*, and *Articles with inconsistent citation formats* are considered non-semantic categories. These categories are excluded in the semantic representation of the Wikipedia concepts. Wikipedia also has across-lingual (*interlanguage links*) feature, which is used to link an article in one language to corresponding articles in other languages. In this paper, this feature is used to extract the title of the English concept that maps to the Arabic article.

Since its release in March 2012, the database dump of AWP has been used as a source of knowledge in the alignment between WordNet synsets and Wikipedia articles. Table 2 shows statistics for different components of AWP. Wikipedia contains several non-useful pages that cannot be linked to any WordNet synsets. The non-useful pages have the following types: articles titled by numbers (years, dates, or others), template pages, and disambiguation pages. Therefore, these pages are filtered out in the mapping.

Most Wikipedia concepts can be classified as NEs, which cover personal names, locations, organizations, and media. It seems to be obvious that *linguistic-term* synsets in WordNet cannot be linked to the NE concept in Wikipedia. The motivation for classifying the Wikipedia concepts before applying the mapping is illustrated in Table 3, which presents some examples of the NE articles that are retrieved for the

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

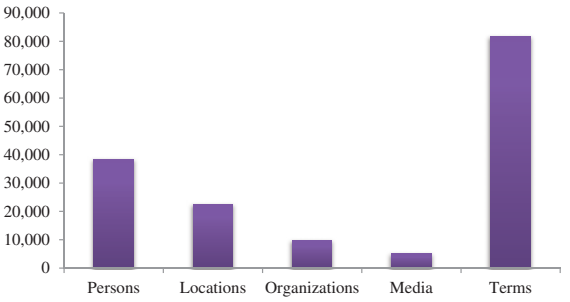


Fig. 1. (Colour online) Distribution of the types of Arabic Wikipedia articles.

linguistic-term synsets. Therefore, the identification of Wikipedia NEs in this study plays an important role in pruning the article candidates for a WordNet synset, thus, decreasing the degree of ambiguity. A rule-based approach (Saif, Ab Aziz and Omar 2013) is used to classify Wikipedia concepts based on the labels of their categories into two types: NE concepts or linguistic-term concepts, as shown in Figure 1. This approach depends on the *keywords* that have been manually identified for each NE type. For instance, the keywords of a NE person contain terms (e.g. أشخاص '*people*') that can be used to form the patterns (e.g. أشخاص_*) of the categories for people, such as أشخاص_على_قيد_الحياة '*Living people'*, and أشخاص_من_أصفهان '*People from Isfahan*'. Thus, when the categories of an article do not match the patterns of NE types, the article is classified as a linguistic term, otherwise it is classified as a NE.

4 Wikipedia-based semantic representation

Semantic representation is a very important topic in cognitive psychology and computational linguistics. It is mainly concerned with studying how the meaning of linguistic units (words, concepts, and texts) can be represented based on the semantic features of the knowledge sources. The semantics of units can be modeled from the associations among the elements in the source of knowledge using the semantic network (e.g. WordNets) or semantic space (unstructured data such as a corpus). The explicit semantic analysis (ESA) approach has been proposed to represent the semantic profile of words in a high-dimensional vector space derived from a structured knowledge of Wikipedia articles by Gabrilovich and Markovitch (2007). The main idea behind the ESA model is that the semantics of the words can be represented by the explicit concepts in the human-organized knowledge sources. An evaluation of this approach indicated that it has achieved promising results in measuring semantic relatedness and document classifications. In this paper, the ESA approach is used to build the semantic representation of Arabic text over Wikipedia articles.

The ESA approach is a corpus-based method that requires preprocessing, such as the removal of stop-words and stemming. Because Arabic is an inflected language with very rich morphological variations, stemming is an important step that produces a more meaningful list of terms by collapsing all variant forms of each word in the

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Semantic vector of the word

شعر

Arabic Concept	meaning	weight
نمو شعر الإنسان	Human hair growth	154.20
الأدب العربي القديم	Old Arabic Literature	141.86
شعر نبطي	Poem	135.70
شعر (تشريح)	hair	132.61
مرض الصلع الوراثي	Androgenic alopecia	129.53
شعر درامي	Dramatical poetry	126.44
جريب الشعرة	Hair follicle	123.36
شعر (أدب)	Poetry	117.19
شعر الدقن	Beard	111.02
إزالة الشعر	Hair removal	104.86

Semantic Vector of the bag-of-words

[شعر, ادب, قصيدة, شاعر, كتابة, لغة]

Arabic Concept	meaning	weight
شعر نبطي	Poem	51.59
أدب	Literature	47.48
الأدب العربي القديم	Old Arabic Literature	47.07
شعر حر	Free verse	42.63
شعر درامي	Dramatical poetry	41.86
شعر (أدب)	Poetry	41.56
شعر عربي	Arabic poetry	39.07
قصيدة	Qasida	37.52
لغة عربية	Arabic language	37.47
شعر المحاورة	Conversational Poetry	37.04

Fig. 2. Examples of top ten concepts in semantic vectors.

corpus together. The stemming of articles also reduces the dimension of the semantic representation matrix. Therefore, a hybrid stemming method (Alhanini and Ab Aziz 2011) is used to extract the stem of each word from the whole of Wikipedia (as plain text).

The ESA approach depends on the distribution of the words in the plain texts of Wikipedia articles when capturing the semantics of the words. To describe ESA in detail, let $W = \{w_1,w_2,\ldots,w_n\}$ be the set of Wikipedia articles and $T = \{t_1,t_2,\ldots,t_m\}$ be the set of terms in the entire Wikipedia after the stemming step. Then, a sparse matrix C is built in which the rows represent the terms in Wikipedia and the columns represent Wikipedia articles. An entry $c[i, j]$ in the matrix is assigned to the *TFIDF* value of term t_i in article w_j. Finally, the semantic profile of a term is represented by a vector of Wikipedia concepts paired with their weights that reflect the association degree of each concept to the term.

The semantic representation of a text is constructed as the centroid of the semantic vectors of its constituent words. The centroid handles the ambiguity for words that have more than one sense in the bag-of-words of the given text. From the perspective of the ambiguity in the language, the semantic vector of an ambiguous word is populated by conflating all the word's senses that appear in different instances in the Wikipedia corpus. In other words, the highest weighted elements in the semantic vector of the ambiguous word may include a conflation of the concepts that are related to the senses of the word. As a consequence, ESA at the word level cannot address the ambiguity in building the semantic representations, especially for words that have totally different senses. For example, the word 'شعر' with the part of speech *noun* is an ambiguous word that has two different senses, 'hair' and 'poetry'. Figure 2 shows the top ten elements with the highest weighting in the semantic vector of this ambiguous word and the centroid vector of the text (شعر, ادب, قصيدة, شاعر, كتابة, لغة) (*poetry, art, Qasida, Poet, writing, language*) that contains this word with five related words to the second sense '*poetry*'.

5 Mapping method

5.1 Candidate extraction

This step extracts suitable articles from Wikipedia as the set of candidates for the given WordNet synset. It depends on a matching between the identifier of this synset

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

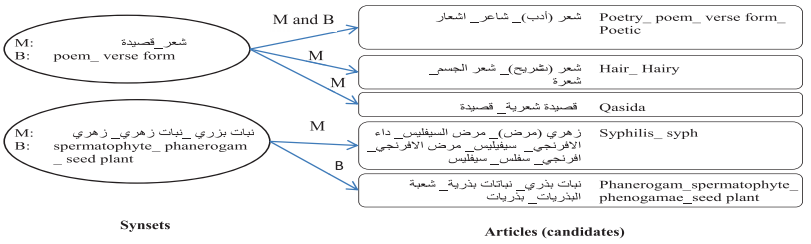


Fig. 3. (Colour online) Example of candidate extraction of synsets.

and the identifier of each Wikipedia article. The identifier of the synset is represented by the monolingual and bilingual features as the set of keyphrases that includes the synonyms in AWN and the synset of the English WordNet that maps to the Arabic synset. The identifier of the article consists of the title and redirect titles (the words or expressions that redirect to the article) as the monolingual features, and the title with the redirect titles of the English article that links to the Arabic article by the interlanguage link as the bilingual features. The AWP database dump contains only the title of the corresponding article in another language; however, the redirect titles of the English article are extracted from the English Wikipedia database dump.

For a synset *syn*, each article in Wikipedia is extracted as a candidate of this synset if at least one element in its identifier exactly matches one element in the identifier of the article. The *description tag* in the title of the article that is used to address the ambiguity is ignored in the matching. Let $W = \{w_1, w_2, \ldots, w_n\}$ be the set of Wikipedia articles. The candidates of the given synset *syn* are defined as the following set:

$$\text{Candidates}(syn) = \{x : x \in W, \text{matching}(iden(syn), iden(x)) > 0\}$$

where the matching function returns the summation of the intersection (the number of elements that occur in two sets) of monolingual and bilingual identifiers between the synset and the article. Consequently, the candidate set for each synset may contain extracted articles based on the matching of monolingual, bilingual, or monolingual–bilingual features. Figure 3 shows examples of candidate extraction for two synsets.

5.2 Measuring semantic similarity

In this paper, we propose three measures that use the monolingual and bilingual features of the synset and article to compute the similarity between them. For the monolingual features, both WordNet and Wikipedia are exploited as the knowledge sources to define the unit-based measures (article-based and synset-based) that model the degree of association between the synset and the article. The unit-based (monolingual) measures are modeled based on Arabic features in two directions (from WordNet to Wikipedia and vice versa). For the bilingual features, the synonyms of English words are used to compute the similarity between the synset and the article. The following subsections describe in detail these three similarity measures.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

5.2.1 Article-based measure

For this measure, Wikipedia is exploited as the source of knowledge to model the semantic relatedness by hybridizing the corpus-based and hyperlink-based methods. The synset is represented as a set of Arabic words extracted from the semantic information in WordNet. After that, the corpus-based technique is used to extract the core semantic units (concepts) of the text representation of a synset. For the bag-of-words set of the synset, the Wikipedia-based semantic representation method (Section 4) is used to construct the semantic vector of this set of words. The non-zero elements in the semantic vector contain the Wikipedia articles that have a semantic association with the synset (an article with higher weight is more related to the synset than another article). Then, the highest scoring elements (ten in this work) from the semantic vector are selected as the core semantic units for the given set of words (synset representation).

For the core semantic units of the synset, the hyperlink-based method is used to measure the semantic relatedness between each unit (concept) and the given article. This hyperlink-based method was inspired by the *normalized Google distance* (Cilibrasi and Vitanyi 2007), which models the semantic distance (SD) between two concepts based on the incoming links as:

$$SD(c_1, c_2) = \frac{\log\left(\text{Max}\left(|IL(c_1)|, |IL(c_2)|\right)\right) - \log\left(|IL(c_1) \cap IL(c_2)|\right)}{\log\left(|N|\right) - \log\left(\text{Min}\left(|IL(c_1)|, |IL(c_2)|\right)\right)} \tag{3}$$

where $IL(c_i)$ is the set of concepts that link to the article c_i. Since the range of SD is between zero and infinity, this measure can be converted to determine the semantic relatedness rather than distance with the following formula:

$$Sim_{hl}(c_1, c_2) = e^{-2 \times SD(c_1, c_2)} \tag{4}$$

Finally, the semantic relatedness between a synset and an article is computed as the average of the similarity between the article as a unit and each concept in the core semantics of the synset as follows:

$$Sim_{WP}(syn, \text{Art}) = \frac{\sum_{c \in CS(syn)} Sim_{hl}(c, \text{Art})}{|CS(syn)|} \tag{5}$$

5.2.2 Synset-based measure

AWN is used as the source of semantics to model the similarity between synsets and articles. An article is represented by its features from Wikipedia (title, redirects, categories, out-links, and page definition) as a set of keyphrases (the preference is for multiword expressions). Then, this text representation of the article is transformed into the set of the synsets (the semantic core of the article) in AWN by adapting the PPR algorithm (Agirre and Soroa 2009). This algorithm has been introduced for word-sense disambiguation or semantic measures, depending on the semantic graph of WordNet. For a set of key phrases, the output of this algorithm is the semantic representation vector of the structural relevance of AWN concepts to the given set of key phrases. This semantic representation vector contains the weight of each node

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

(synset) in the semantic graph according to its semantic relatedness to the given text (a node with high weight is more related to the text than a node with low weight). In the synset-based measure, the top ten highest weighted nominal concepts in the PPR vector are selected as the core semantic units of the given set of key phrases of the article. Therefore, the semantic similarity between the synset and the article is modeled based on the relations between the core semantic units of the article and the synset using the taxonomical relations in WordNet.

The taxonomical relations are exploited to build the WordNet semantic hierarchy, which can be used to measure the semantic similarity between any two synsets. Based on the semantic hierarchy, measures can be classified as path-based measures (Wu and Palmer 1994; Li, Bandar and McLean 2003) or information content measures, using either a corpus-based method (Resnik 1995; Leacock and Chodorow 1998) or an intrinsic-based method (Seco, Veale and Hayes 2004; Zhou, Wang and Gu 2008). In this paper, we use an intrinsic-based measure (Sánchez, Batet and Isern 2011) where the specificity of a concept quantifies the amount of information it provides for different features that relate to the position of the concept in the semantic taxonomy. This measure of the information content of a concept is based on the leaves and the subsumers of the concept, which correspond to the degree of its generality and concreteness in the semantic taxonomy, respectively:

$$IC_{\text{Sanchez}}(c) = -\log\left(\frac{\frac{|\text{Leaves(c)}|}{|\text{subsumers}(c)|}+1)}{\text{Max}_{\text{Leaves}}+1}\right) \tag{6}$$

where $\text{Max}_{\text{Leaves}}$ is the number of leaves in the taxonomy. According to Pirró and Euzenat (2010), the semantic similarity between two nodes (synsets) in a taxonomy is measured by taking into account the information content of the least common subsumer of the nodes (*lcs*) as:

$$Sim_{IC}(c_1, c_2) = \frac{IC_{\text{Sanchez}}(lcs(c_1, c_2))}{IC_{\text{Sanchez}}(c_1) + IC_{\text{Sanchez}}(c_2) - IC_{\text{Sanchez}}(lcs(c_1, c_2))} \tag{7}$$

Finally, the synset-based measure between the article and synset is computed according to the information content similarity of the synset and each concept in the core semantics of the article as:

$$Sim_{WN}(\text{Art}, syn) = \frac{\sum_{u \in CS(\text{Art})} Sim_{IC}(syn, u)}{|CS(\text{Art})|} \tag{8}$$

5.2.3 Bilingual-based measure

The bilingual-based measure is introduced to exploit the English words in WordNet and Wikipedia that correspond to the Arabic synset and article using the cross-lingual links. This measure focuses only on the most important relation (synonyms) in English WordNet and the corresponding relation in Wikipedia (redirects) to compute the similarity based on the bilingual features. Therefore, the synset and article are represented as the set of key phrases of the English synonyms that correspond to the given synset and the title of the English article with its redirect titles that link to the Arabic article by the interlanguage link. To determine the

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

degree of the importance of the elements in the bilingual representation sets for the synset and article, each lemma in the two sets is weighted based on the number of noun senses of the lemma in WordNet and the number of articles that contain this lemma in their English identifiers as:

$$\text{Weight (lemma)} = \begin{cases} \frac{1}{|\text{Senses}_{\text{noun}}(\text{lemma})|} \text{where, lemma} \in \text{synset} \\ \frac{1}{|\text{Articles(lemma)}|} \text{where, lemma} \in \text{article} \end{cases} \quad (9)$$

Finally, the bilingual measure is used to find the similarity between the synset and the article based on the weight of the lemmas in the intersection between the bilingual features of the synset and the article. This measure is normalized to the maximum length of the sets:

$$Sim_{BF}(Syn, \text{Art}) = \frac{\sum_{w' \in BF(Syn) \cap BF(\text{Art})} \text{weight}\left(w'_{\text{synset}}\right) \times \text{weight}\left(w'_{\text{Article}}\right)}{\max\left(|BF(Syn)|, |BF(\text{Art})|\right)} \quad (10)$$

5.3 Matching classification

Starting with AWN synsets, this step classifies each synset in AWN as *mapped* or *not-mapped*. The mapped class is when the synset has an identical concept in Wikipedia. The *not-mapped* class means the synset cannot be matched to any articles or the semantic associations between the synset and the matched article are very limited. Based on the candidate extraction step for each synset, there are two possible cases for the synset in Wikipedia: *null-matched* (no articles can be retrieved for the synset) and *matched* (at least one article matches the synset). The synsets that are null-matched are classified directly as *not-mapped* to all Wikipedia articles. For the synsets that have candidates in AWP, the semantic similarities are used to decide whether the synset can be linked to an article in Wikipedia.

To classify the matching, let $Cand(syn) = \{a_1, a_2, \ldots, a_m\}$ be the set of Wikipedia articles (candidates) for a given synset *syn*. Three semantic similarity vectors are constructed by measuring the similarity between the synset and each article in the candidate set using three measures: article-based, synset-based, and bilingual measures. Each vector contains the similarity scores of the given synset and its candidates obtained for one similarity measure. For instance, the similarity vector of the article-based measure (Sim_{WP}) is defined as: $SV_{WP} = \{a_1, v_{wp1}, a_2, v_{wp2}, \ldots, a_m, v_{wpm}\}$ where v_{wpi} refers to the similarity score between *syn* and the $a_i(Sim_{WP}(syn, a_1))$. Due to the differences between the granularity of the semantic contents in WordNet and Wikipedia, the matching classification is introduced to handle two important issues that can occur in the matching of identifiers for the synset. These issues reflect the variety of semantic associations between the synset and its candidates. In non-similar semantics, the semantics of all candidates are totally different from the synset, whereas in multi-similar semantics, the semantics of the synset are related to more than one candidate. In non-similar semantic synsets, the selection of the article with the highest similarity score among candidates (as in word-sense disambiguation) is not suitable. On the other hand, the challenge for a synset with

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

multi-similar semantics is how to determine the most related article from the set of candidates.

To handle the non-similar semantic issue, the threshold-based classification technique (Niemann and Gurevych 2011) is used to detect unrelated articles (the candidates with similarity scores below the threshold). The threshold is determined for each similarity measure by searching for the optimum value from zero to one that maximizes the performance (in terms of the F-measure) of the mapping for the Arabic gold standard data set (Section 7.1). Based on the threshold of each similarity measure, its corresponding semantic similarity vector is used separately to assign the binary weight with respect to the similarity score for each article in the candidate set: 1 (identical or related to the synset) or 0 (unrelated). For the synset–article pair, syn–a_i is assigned to 1 if the similarity score v_{wpi}exceeds the threshold. Consequently, the binary weights of the synset–candidate pairs using three semantic similarity vectors can be combined together as:

	SV_{wp}	SSV_{wn}	SSV_{bf}	Combined
$syn - a_1$	x_1	y_1	z_1	c_1
$syn - a_2$	x_2	y_2	z_2	c_2
.	.	.	.	.
.	.	.	.	.
$syn - a_m$	x_m	y_m	z_m	c_m

where x_i,y_i, and z_i are the binary weights (0 or 1) that have been assigned to the $syn - a_i$ pair depending on the semantic vectors of the article-based, synset-based, and bilingual-based measures, respectively. The c_i is the combined binary weight (0 or 1), which depends on the majority voting function,[3] $c_i = x_i y_i + x_i z_i + y_i z_i$, of three inputs ($x_i$, y_i, and z_i). This combined weight of all pairs means that the synset can be semantically related to none, one, or more than one article (multi-similar semantics issue). For a synset with a combined weight of 0 for all candidates, the synset–article pairs are classified as *not-mapped* (none-similar articles for this synset). For a synset that relates to only one article a_i, the syn–a_i pair only is classified as mapped, and the other pairs are classified as *not-mapped*.

Since the aim of the proposed method is to map a synset to only one Wikipedia article, therefore, for each synset, the mapped class must be assigned to only one synset–article pair. From the candidates, the most related article is selected according to the highest similarity score of the integrated measure (the total score of the three measures). In spite of the similarity scores of the synset and article being ranged between zero and one for the monolingual measures, however, these scores are not balanced across any two measures because each measure models the similarity depending on different sources for the semantics. For this reason, the similarity vectors of monolingual measures (article-based and synset-based) are normalized

[3] This function, $c_i = ((x_i \text{and} y_i)\text{or}(x_i \text{and} z_i)\text{or}(y_i \text{and} z_i))$, returns 1 (true) if at least two variables are 1.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 4. *Results of mapping for linguistic terms*

# Extracted candidates per synset	# Synsets	# Mapped synsets	Percentage of mapped
1	2,295	2,206	96.12 per cent
2	1,802	1,637	90.84 per cent
3	843	786	93.24 per cent
4	397	348	87.66 per cent
>4	482	403	83.61 per cent
No candidates	994	–	–
Total	6,813	5,380	78.97 per cent

to balance the similarity scores in two vectors. In each vector, the normalization function for each score depends on the maximum and minimum values in the vector as:

$$\text{norm}\,(v_i \in SV) = \frac{v_i - \min\,(\text{SV})}{\max\,(\text{SV})} \tag{11}$$

Consequently, the synset–article pair where the article has the highest integrated similarity score is classified as a *mapped* label. Other synset–article pairs are classified as *not-mapped*, regardless of their similarity scores.

6 Full mapping results

This section reports the results that have been obtained by applying the proposed method on the synsets in AWN to AWP articles. Since the concepts in Wikipedia are nominal, the synset–article mapping method was carried out only on the AWN nominal synsets. AWN contains 7,960 nominal synsets, which were classified manually by lexicographers into NE synsets (1,147) and linguistic-term synsets (6,813). On the other hand, the Wikipedia concepts were classified into two main types, linguistic terms and NEs (including personal names, locations, organizations, and movie names), using the rule-based approach (as mentioned in Section 3.2). For the linguistic terms, Table 4 presents the mapping method results for the synsets according to the number of articles that can be extracted for each synset group.

For the NE synsets, the mapping was applied only to 334 entities that have corresponding synsets in PWN through the bilingual relation *equivalence*. Other NE synsets were excluded from this study because the mapping depends on the bilingual and monolingual features in both candidate extraction and measuring the similarity. The results of mapping these NE synsets to NE articles are shown in Table 5. In the candidate extraction step, NE synsets were classified based on the number of articles that could be retrieved for each synset into monosemous (one article only) and polysemous (more than one article).

From the mapping results, it is clear that the ratio (96.41 per cent) of mapped NE synsets is higher than the ratio (78.97 per cent) of mapped synsets of linguistic terms. This significant difference can be explained by the following factors that can affect the mapping: the compatibility of terming the units in two sources for each type, the coverage of knowledge sources (number of synsets for which candidates

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 5. *Results of mapping for NEs*

# NE synsets	Type of synsets	# Mapped synsets	Percentage of mapped
Monosemous	188	185	98.40 per cent
Polysemous	146	137	93.84 per cent
All	334	322	96.41 per cent

Table 6. *Coverage of AWN against AWP*

		Single words			Multiword		
Sources	Type	Monosemous	Polysemous	All	Monosemous	Polysemous	All
AWN	NE	652	163	815	628	18	646
	LT	3,470	2,156	5,626	3,321	147	3,468
	All	4,122	2,319	6,441	3,949	165	4,114
AWP	NE	21,844	1,232	23,076	94,896	784	95,680
	LT	30,444	2,029	32,473	98,440	1,269	99,709
	All	52,288	3,261	55,549	1,93,336	2,053	1,95,389

can be extracted from AWP), and the degree of ambiguity for each type. Since NEs include the distinct entities that have the same names in the different sources, such as person and locations, the terming compatibility and unambiguity degree of NEs in AWN and AWP are indeed better than linguistic terms, which can be represented in different ways. For the linguistic terms, there are 6,813 synsets, and the candidate extraction step retrieved possible articles for only 5,819 synsets, but all NE synsets were covered by Wikipedia.

Based on the coverage of knowledge sources, we compared the number of lemmas (single or multiword) for AWN and AWP that are covered by the identifiers of each source (synonyms in AWN and a title with redirects in AWP). Table 6 presents the statistics of monosemous and polysemous lemmas for the single and multiword lemmas in both sources. From this table, the coverage of AWP, which contains 2,50,938 lemmas, is much higher than AWN, which includes only 10,555 lemmas (single and multiword).

7 Evaluation

7.1 Test set

To evaluate the proposed synset–article mapping, a gold standard data set that contains the WordNet synsets and their corresponding Wikipedia articles is required. To our knowledge, there is no gold standard (available evaluation data sets) for the alignment of WordNet synsets and Wikipedia articles in Arabic. Therefore, we created an Arabic data set for evaluating the proposed method in addressing the ambiguity and synonymy issues in the alignment of WordNet synsets and articles in Wikipedia.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 7. *Statistics for the mapping data set*

Samples	Synsets	Articles	Avg. of candidates	Mapped	Reliability
1	188	692	3.68	129	0.94
2	30	115	3.83	30	0.99
3	200	484	2.42	193	0.98
All	418	1,291	3.09	352	0.97

Words in AWN can be classified into two types: monosemous (the word has only one sense) and ambiguous words (each word has more than one sense). Since the proposed method has been introduced to address the ambiguity issue in mapping synsets to Wikipedia articles, the data set was created to encompass the ambiguous words in WordNet and Wikipedia. From WordNet, we randomly selected 50 ambiguous words, which returned 188 synsets with the part of speech *noun*, as the first sample. From Wikipedia, 115 articles that could be candidates for 30 synsets by matching their identifiers were selected as the second sample. The synsets in these two samples were selected to cover the ambiguous single words in the two sources, WordNet and Wikipedia. The synsets in the final sample were selected from WordNet using two factors: the degree of specificity of the synset and whether the extracted candidates had monolingual or bilingual features. To ensure this sample covers various degrees of specificity, 200 synsets are selected from different positions (based on the depth of the synset) within a semantic taxonomy of AWN. These selected synsets were distributed according to their candidates from Wikipedia to include the following cases: monolingual only (all candidates of a synset were extracted by matching Arabic features only), bilingual only, and monolingual–bilingual together (the candidates were retrieved by matching both Arabic and English features).

The manual mapping annotation for the compiled data set was accomplished by three human annotators (experts in computational linguistics). The semantic information of the synset (synonyms, related words, and gloss) and the article (title, category, and definition) were presented to provide the annotators with sufficient information to make a decision. For each synset–article pair, the annotators were asked to classify this pair as one of the following categories: mapped (1) or *not-mapped* (0). The mapped category is when the WordNet synset is identical or related to a large extent to the Wikipedia concept in the synset–article pair. The not-mapped category indicates the semantics of the synset are completely different from the semantics of the article or the relation between them is very limited. Finally, the synset–article pair was assigned to a category (*mapped* or *not-mapped*) if at least two annotators marked the pair as this category. Table 7 presents the statistics of the Arabic mapping data set.

7.2 Performance measures

The proposed method was evaluated for the manual data set by comparing the category obtained by the method to the category assigned by the human experts for

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 8. *Evaluation results for different representations of synsets and articles*

Representation	WP features	WN features	P	R	F	Accuracy
R11	T+C	S+NH	0.707	0.820	0.759	0.855
R12	T+C	S+G	0.762	0.869	0.812	0.888
R13	T+C	S+NH+G	0.779	0.889	0.830	0.899
R21	T+L	S+NH	0.775	0.906	0.835	0.901
R22	T+L	S+G	0.792	0.926	0.854	0.912
R23	T+L	S+NH+G	0.802	0.937	0.864	0.918
R31	T+D	S+NH	0.765	0.894	0.825	0.894
R32	T+D	S+G	0.782	0.914	0.843	0.905
R33	T+D	S+NH+G	0.790	0.923	0.851	0.910
R41	T+C+L	S+NH	0.775	0.906	0.835	0.901
R42	T+C+L	S+G	0.792	0.926	0.854	0.912
R43	T+C+L	S+NH+G	0.807	0.943	0.870	0.921
R51	T+C+D	S+NH	0.778	0.909	0.838	0.902
R52	T+C+D	S+G	0.797	0.931	0.859	0.915
R53	T+C+D	S+NH+G	0.812	0.949	0.875	0.924
R61	T+D+L	S+NH	0.797	0.931	0.859	0.915
R62	T+D+L	S+G	0.819	0.957	0.883	0.929
R63	T+D+L	S+NH+G	0.826	0.966	0.891	0.934
R71	T+C+D+L	S+NH	0.792	0.926	0.854	0.912
R72	T+C+D+L	S+G	0.817	0.954	0.880	0.928
R73	T+C+D+L	S+NH+G	0.829	0.969	0.893	0.936

each synset–article pair in terms of the performance measures (precision, recall, F-measure, and accuracy). The accuracy measure is used to determine the percentage of correctly identified mapped and not-mapped classes compared to the manual classifications in the data set. Precision (P) denotes the ratio of correctly mapped to all mapped classes assigned by the method. Recall (R) determines the ratio of correctly mapped to the manually mapped class in the data set. The F-measure is the harmonic mean of the precision and recall.

7.3 Experiments

7.3.1 Representation features

This experiment aimed to examine the performance of the method with different representations of the monolingual features from AWN and AWP. The text representation of a synset can be extracted as the bag-of-words set from the following semantic information in AWN: synonyms (S), (NH) relationsand glosses (G). On the other hand, the text representation of an article can be extracted from the following Wikipedia features: titles (T) (title of article with redirect titles), categories (C), out-links (L), and the page definition (D). Table 8 presents the evaluation results for different representations of articles and synsets for the data set, which shows that feature representations of the synsets and articles play an important role in the performance of the mapping method. The accuracy for the data set ranged between 85.5 per cent and 93.6 per cent for the different representations of synsets and

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

articles. The representation R73, which combines all semantic features for two units, had the highest accuracy (93.6 per cent) in comparison to other representations. For the articles, the features can be ordered based on their degree of importance as: out-links (L), definitions (D), and categories (C). On the other hand, in addition to the synonyms, the gloss feature of the synset improved the performance of the method better than the NH feature.

7.3.2 Unit-based versus text-based methods

For Arabic features, the semantic similarity between a synset and article was computed using the unit-based measures (synset-based and article-based) that depend on the PPR and ESA semantic representation vectors. In these measures, the PPR and ESA representation vectors were pruned by selecting the top ten weighted concepts from each vector as the semantic core concepts of the synset and article. To investigate the effect of this pruning, the unit-based measures are compared with the text-based measures that use the direct semantic representation vectors of the PPR algorithm and ESA model. The latter measures are classified as text-based measures because the similarity between the synset and the article is computed as the similarity between two text fragments of the synset and article. Based on the semantics in AWN, the first similarity measure is defined by the cosine function between the PPR representation vectors of the synset and article as:

$$TB_{WN}(\text{Syn}, \text{Art}) = \frac{\text{PPR}(\text{syn}) \times \text{PPR}(\text{Art})}{|PPR(syn)|\,|PPR(Art)|} \tag{12}$$

The second measure uses the semantics in Wikipedia to compute the similarity between the synset and article. For two bag-of-words sets, the ESA vector is constructed for each set. Then, the similarity measure between the two vectors is defined by the Jaccard function as:

$$TB_{WP}(Syn, Art) = \frac{ESA(Syn) \cdot ESA(Art)}{|ESA(Syn)|^2\,|ESA(Art)|^2 - ESA(Syn) \cdot ESA(Art)} \tag{13}$$

To evaluate the text-based technique (TB), the synset-based and article-based measures are replaced by the text-based TB_{WN} and TB_{WP} measures in the proposed mapping. For the evaluation data set, Figure 4 shows the accuracy of the mapping using the TB for the similarity measure (TB_{WN} and TB_{WP}) compared to the unit-based technique for the similarity measures (synset-based and article-based) for the same representations of the synset and article. The comparison shows that the unit-based technique is more accurate than the TB. In the unit-based technique, the TB (building the semantic vectors) was used only to construct the semantic core of the synset in AWP and the semantic core of the article in AWN (as a feature selection task). The hyperlinking feature in Wikipedia and the specificity of the concepts were exploited in the unit-based technique to gather more semantic evidence than was possible for the TB in modeling the similarity.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

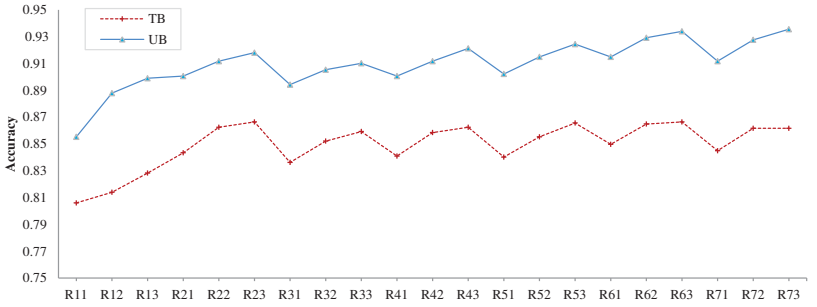
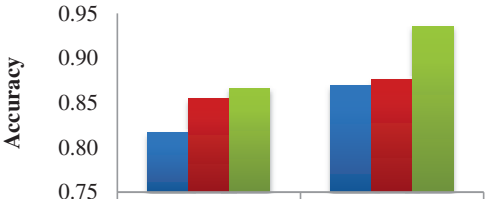


Fig. 4. (Colour online) Comparison between text-based (TB) and unit-based (UB) techniques.



	Text-Based	Unit-Based
WN	0.816	0.869
WP	0.855	0.876
WN-WP	0.866	0.936

Fig. 5. (Colour online) Comparison of AWN, AWP, and WordNet–Wikipedia (combined) for semantic measures.

7.3.3 *Semantics in AWN compared to AWP*

In addition to the bilingual measure, the proposed method integrated the synset-based with the article-based measures that depend on the semantics in AWN and AWP. To investigate the effects of integrating the semantics for two different sources, the mapping is evaluated by exploiting only one monolingual similarity measure[4] that reflects the efficiency of the corresponding source (AWN or AWP). Figure 5 shows the best accuracy of the mapping using the semantic similarity of monolingual features based on the semantics in the following sources: only AWN, only AWP, and a combination of them. The evaluation shows that the performance of the mapping based on the combined similarity of the two sources is better than using an individual measure for either the unit-based or TB. For a single monolingual source, the similarity based on the semantics of AWP outperformed the semantics of AWN. This is due to the limitations of AWN in terms of its coverage of words and the lack of semantic and lexical relations.

7.3.4 *Effects of bilingual features*

To determine the effect of bilingual features on the mapping, the monolingual method is assessed for the same data set. The monolingual method is a mapping method using

[4] The similarity between a synset and an article is computed by integrating the bilingual-based measure with either only the synset-based measure or only the article-based measure.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

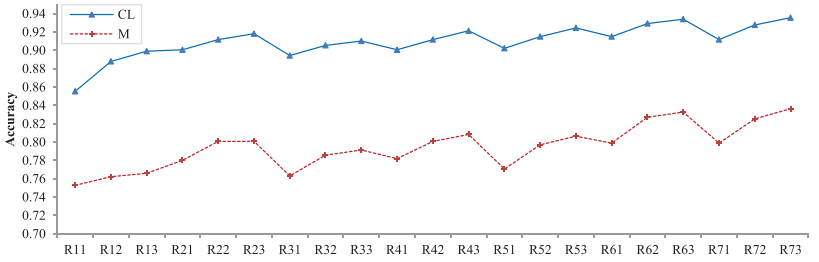


Fig. 6. (Colour online) Monolingual (M) features versus cross-lingual (CL) features for mapping.

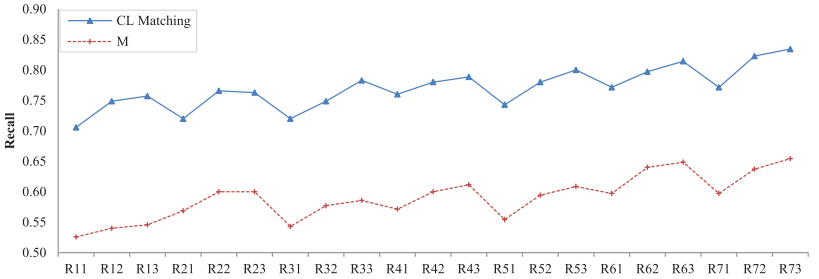


Fig. 7. (Colour online) Monolingual (M) features versus cross-lingual (CL) features for only candidate extraction.

only Arabic features for extracting candidates and measuring semantic similarity. In the matching classification step of the monolingual method, a candidate is classified as *mapped* or *not-mapped* based on only two monolingual measures (synset-based and article-based). Figure 6 shows a comparison between the monolingual method (M) and a method that combines Arabic features with English features ($M \cup B$). From this figure, it can be observed that the proposed method that exploited the monolingual and bilingual features had better accuracy than the monolingual method.

Bilingual features in candidate extraction To examine the matching for the mapping data set, the monolingual method is compared to the same method but using bilingual features for candidate extraction only, as shown in Figure 7. The performance of the method is evaluated using the recall (R), which reflects the ratio of correctly mapped articles compared to the mapped class in the data set. For each synset in AWN, the ambitious goal of this step is that for a synset, the identical concept in AWP should be included in the extracted candidate set. The aim of using the bilingual features in this step is to handle the incompatibility of the concept's identifiers in AWN and AWP. The evaluation results in Figure 7 show that the chance of including the correct concept in the candidate set for some synsets increased by using the bilingual features. In other words, the cross-lingual matching was better than monolingual matching for mapping.

To determine the effect of the bilingual features in candidate extraction, we evaluated this matching step for the mapping data set in terms of the capability of extracting the corresponding concepts of the synsets. Thus, only the synset–article pairs that have been annotated by the experts as a mapped class are selected (129,

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 9. *Monolingual versus bilingual matching for352 mapped pairs*

	M	B	M ∩ B	B-(M∩B)	M-(M∩B)
Statistics	250	312	210	102	40
Average	0.71	0.89	0.60	0.29	0.11

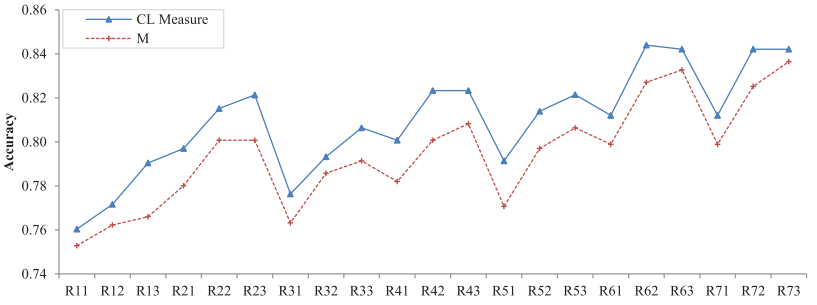


Fig. 8. (Colour online) Monolingual (M) features versus cross-lingual (CL) features for measuring only similarity.

30, and 193 pairs for three samples of the data set). For the selected pairs from the data set, let M and B be the sets of identical articles that have been extracted using the monolingual and bilingual features, respectively. Then $B - (M \cap B)$ is the set of identical articles that have been retrieved by matching the bilingual features and could not be extracted using the monolingual feature. This set reflects the contribution of the bilingual features for handling the incompatibility of terming the concepts in the two sources for the manually mapped pairs. For the mapped synset–article pairs, Table 9 presents the statistics of correct articles that have been retrieved by matching the monolingual and bilingual features.

Bilingual features in semantic similarity In measuring the semantic similarity, the important semantic relations (synonyms) were used to model the similarity based on the bilingual features. Figure 8 shows the accuracy of the monolingual measure and the integrated cross-lingual similarity measure (monolingual and bilingual) for the data set.

In summary, the evaluation of the monolingual method is considered to be the baseline. The effect of bilingual features on the performance of the mapping can be determined as the improvement ratio of the baseline results against the results when incorporating bilingual features in candidate extraction (Cand), measuring semantic similarity (Sim), and both (Cand& Sim). Figure 9 shows the improvement ratio for the performance method when using bilingual features against the baseline results.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 10. *Characteristics of the state-of-the-art mapping methods adapted for Arabic*

Methods	Proposed by	Similarity			WP features	Mapping decision
		Types	Technique	Measure		
RC05	Ruiz-Casado *et al.* (2005)	TB	VSM with TFIDF	Cos	Full article	Max
PN10	Ponzetto and Navigli (2010)	TB	Gloss-based	Overlap	T, L, C	Max
NG11	Niemann and Gurevych (2011)	SB	PPR	X^2	T, D, C	Threshold
FS12	Fernando and Stevenson (2012)	TB	Gloss-based	Normalized Overlap	Full article	Threshold
NP12	Navigli and Ponzetto (2012)	SB	Graph-based	Path-based	T, L, C	Max

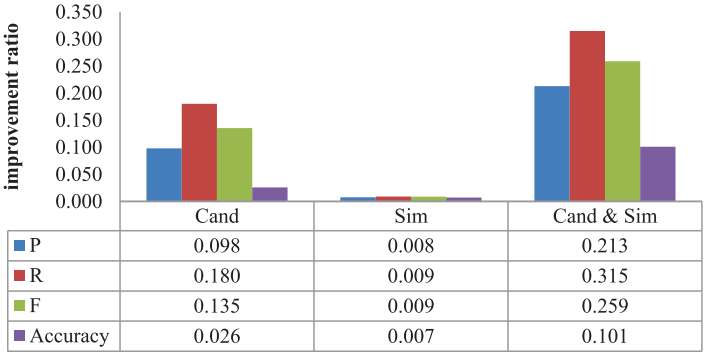


Fig. 9. (Colour online) Improvement ratio of the accuracy with bilingual features.

7.3.5 Comparison with previous works

From the literature, we adapt five methods[5] for Arabic sources for mapping Arabic synsets to their corresponding articles in Wikipedia using only Arabic features in the matching (candidate extraction) and measuring the similarity. The differences among these methods can be summarized according to the following characteristics (as in Table 10): type of similarity method (text-based or structure-based), semantic representation technique, similarity measure, representation features, and the mapping criteria (maximum similarity or threshold). Since AWN does not have glosses, we use all semantic information (S, NH, and G) in AWN to represent the synset in all methods.

[5] We re-implemented these five methods to make them applicable to Arabic sources.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

For evaluation, these monolingual-based methods are applied to the synsets in the Arabic data set to compare the mappings by different methods against the manually labeled ground-truth mappings. In all methods, only the Arabic identifiers of the synsets and articles are used to retrieve the candidates (articles) for each synset based on the matching of the lemmas. It must be pointed out that the candidates in the data set that were extracted using only bilingual features have been excluded when measuring the similarity and the threshold classification. In other words, a synset with a bilingual candidate pair is classified directly as *not-mapped* if the Arabic identifiers of the synset are different from the identifiers of the article. In the NG11 and FS12 methods, a threshold-based classification is used to determine the parameter (similarity threshold) that achieves the highest F-measure based on the manual mappings. For the matching classification, the candidate with the highest similarity value (and which exceeds the threshold in NG11 and FS12) is classified as *mapped*, and other candidates are classified as *not-mapped*. The details of the implementation of measuring the similarity in each method are presented below.

In RC05, the similarity between a synset and an article is computed based on the vector space model, which represents the synset and article as weighted vectors. The semantic information of all nominal synsets in AWN is transformed as the collection of documents, where each document contains the sequence of words from the S, NH, and G features[6] of the corresponding nominal synset. Based on this collection of texts for all nominal synsets in AWN, each synset is represented by the vector that contains the *TFIDF* weights of each word in its representation features. In AWP, the representation vector of an article contains words paired with their frequency in the free text (after stemming and removal of stop-words). Then, the similarity between the synset and article is computed using the cosine measure[7] between their representation vectors. For the given synset and its candidates, the candidate with highest similarity is classified as *mapped*, and other candidates are classified as *not-mapped*.

In PN10 and FS12, a synset and article are each represented as a set of words (disambiguation context) from the features in AWN and AWP. Then, the gloss-based technique is used to measure the similarity between the synset and article-based on their disambiguation context sets. Let $DC(s)$ and $DC(a)$ be disambiguation context sets of the synset$s$ and article a. The similarity in the PN10 method is measured as the intersection between two sets $(|DC(s) \cap DC(a)| + 1)$. It is normalized using the joint probability $p(a, s)$. In the FS12 method, the similarity is computed as

$$Sim_{FS12}(s, a) = \frac{|DC(s) \cap DC(a)|}{\min\left(|DC(s)|, |DC(a)|\right)} \tag{14}$$

For the methods that depend on the structure-based similarity, the semantic graph of AWN is constructed, in which the nodes represent the synsets and the edges represents all relations among synsets. To measure the semantic relatedness for the

[6] In the original algorithm, the synset is represented by the words in the gloss and hypernyms.
[7] The experimental results showed that the cosine measure achieved better accuracy than the dot product.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

NG11 method, the semantic graph of AWN is represented by the row-stochastic transition matrix $M \in \mathbb{R}^{n \times n}$, where the entry M_{ij} corresponds to the probability of moving from node i to j: this is $(1/\text{degree}(i))$ if there exists an edge from i to j and 0 otherwise. The PPR algorithm is used to generate the semantic representation vector of a node or set of nodes by solving the following linear system:

$$P = \alpha MP + (1 - \alpha)\, v \tag{15}$$

where α is the damping factor, which is set to 0.85, and v is an n-dimensional personalization vector whose initialization differentiates between direct PPR and basic PPR. Direct PPR can be used only to generate the representation vectors of synsets; for a given synset, the personalization vector is initialized by assigning the weight of 1 to the element that corresponds to this synset, and the weight of 0 to the others. The personalization vector in basic PPR is initialized depending on the textual representation of the synsets or articles. For a given bag-of-words, let $C = \{c_1, c_2, \ldots, c_k\}$ be the set of synsets (candidates) of the words. The vector is initialized by assigning the weight of $1/k$ to the elements that correspond to the synsets in C, and the weight of 0 to the others. Since the output of the PPR algorithm is the representation vector over the synsets in AWN, for a given synset–article pair, the semantic relatedness is measured as the similarity between their representation vectors using the x^2 measure.

To measure the similarity between a synset and an article in the NP12 method, the semantic graph of AWN is used to build the subgraph (DG) of the disambiguation context of the article. For a given disambiguation context (as a set of words) of article a and synset$s$, the synsets that can be candidates for each word in this set are pooled together with s to represent the nodes of the subgraph. Then, for each node v in DG, the depth-first search algorithm is used for the AWN graph to find the paths of maximal length L between v and $v \in V(DG) - v$, and add the nodes and edges in each path to DG.[8] Finally, the semantic similarity$Sim_{NP12}(s, a)$ is measured based on the lengths of paths between the synset$s$ and each node in the subgraph of the article a as

$$Sim_{NP12}(s, a) = \sum_{s' \in DG(a)} \sum_{p \in Paths(s,s')} e^{-(\text{Length}(p)-1)} \tag{16}$$

where the Length(p) returns the number of edges in the path p. This similarity is normalized using the joint probability as

$$p(a, s) = \frac{Sim_{NP12}(s, a)}{\sum_{\substack{s' \in senseWN(s) \\ a' \in senseWP(s)}} Sim_{NP12}(s', a')} \tag{17}$$

For comparison, Table 11 presents the best evaluation results that have been obtained by applying the mapping methods to the Arabic data set. For theNG11 method, we report the evaluation results for the two versions of the PPR algorithm

[8] We evaluated this method using 2, 3, and 4 as the maximum depth of the depth-first search. The best results were obtained by setting the maximum depth to 4.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 11. *Evaluation of mapping methods for the Arabic data set*

Method	P	R	F	Accuracy
RC05	0.478	0.509	0.493	0.741
PN10	0.485	0.514	0.499	0.745
NG11-B	0.556	0.591	0.573	0.795
NG11-B	0.573	0.609	0.590	0.806
FS12	0.530	0.563	0.546	0.776
NP12	0.507	0.537	0.521	0.760
Proposed-M	0.616	0.654	0.634	0.835
Proposed-CL	0.829	0.969	0.893	0.936

Table 12. *Comparison of the similarity measures for the Arabic data set*

Method	P	R	F	Accuracy
RC05	0.575	0.671	0.619	0.770
PN10	0.577	0.674	0.622	0.772
NG11-B	0.636	0.734	0.682	0.809
NG11-D	0.663	0.766	0.711	0.827
FS12	0.560	0.654	0.603	0.761
NP12	0.619	0.714	0.663	0.798
Proposed	0.829	0.969	0.893	0.936

(basic and direct) for building the semantic vectors to measure the similarity. The difference between these versions is the initialization of the distribution vector, which depends either on the text representation of the synset (TB) or on the synset only as the unit (unit-based technique) for the basic and direct versions, respectively. The evaluation results of the proposed method for two cases (monolingual and cross-lingual features) are shown in the last two lines of Table 11.

For the ambiguity issue, the performance of the mapping methods depends heavily on the technique for measuring the similarity, which is used to determine the degree of the semantic association between the synset and the article. To ensure a fair comparison, we evaluate the similarity measures of the state-of-the-art methods for the Arabic mapping data set using the same experimental settings (representation features and the matching classification). For each synset in the data set, the similarity is measured between this synset and each article that was retrieved based on the monolingual and bilingual features. Since the performance of the similarity measures may be affected by the representation features employed for the two sources, we use the same features (R73) from AWN and AWP to represent the synset and article for each measure. Using the same threshold-based classification technique, the threshold is determined for each similarity measure based on the similarity scores of the corresponding measure. The evaluation results of the similarity measures are presented in Table 12.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 13. *Examples of concepts that have different monolingual identifiers in AWN and AWP*

Arabic synsets		Corresponding articles	
فرس;الفرسيات	equine; equid	الخيليات	Equidae; Equid
مرفاع	Crane	رافعة (آلة)	Crane_(machine)
مواد استهلاكية	Consumer goods	سلع نهائية	Final good; Consumer goods
قنبلة سيارة	Car bomb	سيارة مفخخة	Car bomb
حامل أسهم_صاحب أسهم	shareholder;	مساهم	shareholder
هراوة_مضرب الكرة	Golf club	ميجار	Golf club
باتيلة	Paella	بايلا	Paella
عقيق أحمر	Garnet	جارنيت	Garnet
هاردون	Hadron	هادرون	Hadron
أكل دسم	Junk food	طعام خردة	Junk food

7.4 Discussion

As shown in Table 11, the proposed method for the Arabic mapping task significantly outperformed the state-of-the-art methods that have been introduced for richer-knowledge sources. The accuracy of the proposed method (93.6 per cent) was higher than for the other methods, which ranged from 74.1 per cent to 80.6 per cent. The superior performance of the proposed method can be interpreted as due to the extension of using semantic information in matching and measuring similarity. For handling the lack of semantics in the mapping sources, the semantic evidence in the proposed method was extended by exploiting the cross-lingual features and integrating the semantics of the two knowledge sources. For the cross-lingual extension, the comparison shows that exploiting both monolingual and bilingual features together with the mapping of Arabic sources provides more semantic evidence than using only monolingual features. A comprehensive analysis of the contributions of bilingual features in the mapping was presented in the last section. For the Arabic data that contains 352 synsets that were manually linked to their corresponding articles by linguists, the monolingual-based matching from the previous works failed to retrieve the identical articles for 102 (29 per cent) synsets due to the synonymy issue. Therefore, the recall measure did not exceed 71 per cent in all experiments that used only Arabic for candidate extraction. Table 13 shows some concepts in AWN and AWP that could not be linked using the monolingual-based matching.

The evaluation in Table 12 can be analyzed to compare the semantic similarity measures in the mapping task. For the TB, the similarity measures (RC05, PN10, and FS12) performed poorly for the Arabic data set. For these measures, which depend on the number of words in common, there is no significant difference among the accuracy values (77 per cent, 77.2 per cent, and 76.1 per cent) of the measures using the same representation features. For the WordNet-based technique (NG11-B, NG11-D, and NP12), which depends on the semantic graph of WordNet, the direct PPR algorithm achieved a higher accuracy (82.7 per cent) compared to basic PPR (80.9 per cent) and NP12 (79.8 per cent). Compared to the TB, the WordNet-based technique performed better at mapping, as the accuracy increased from 77.2 per cent

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

(the best accuracy was for the text-overlap technique) to 82.7 per cent, 80.9 per cent, and 79.8 per cent. This improvement is very predictable since the WordNet-based technique can handle the ambiguity and synonymy issues of the words when measuring the semantic similarity. For theNG11 method, measuring similarity based on the direct PPR version, which treats a synset as a unit, outperformed measuring similarity based on basic PPR (where a synset is represented as a set of words). This is because the semantic vector of the synset when using the basic PPR version is initialized by assigning a similar weight for each sense of the words in the text representation. For the synset that contains the ambiguous words, the initialization of the probability distribution builds a vector that includes the conflicting different senses of each word.

Although the results obtained by the NG11-D method as the WordNet-based similarity measure are better than the results for the same method in Table 11, however, these results are still worse than those of the proposed similarity technique that integrated three semantic measures depending on the knowledge in AWN, AWP, and the cross-lingual features from English sources. This may be due to the following reasons: the limitations of AWN, the excessive PPR representation for synsets, and the noisy features of some articles.

The similarity measure in the NG11-D method is based on the PPR algorithm, which has been successfully used for mapping English sources. In general, the semantics in AWN are very limited compared to English WordNet in terms of the coverage of lemmas, the availability of the glosses, and the number of semantic relations. Thus, the performance of this method for the Arabic data set is negatively affected by these limitations of AWN.

The similarity between a synset and each candidate was computed according to the weights in their PPR representation vectors. The semantic representation of the PPR algorithm is considered very excessive, since it contains the weights over all synsets in AWN. Thus, for some synsets, there is no large difference in the similarity scores of candidates for a synset. For instance, the similarity scores between the synset ‘نَجِيل’، ‘*lawn*’ and its candidates (‘نجيل’، ‘*Cynodon dactylon*’ and ‘نبات مروج’، ‘*Lawn*’) using the NG11-D measure are 0.1187 and 0.1193, respectively.

For building the semantic vector of a candidate article, the initialization of the personalization vector in the PPR algorithm depends on the words in the features of the article. The features of some articles contain noise words that are not related to the corresponding concepts. For example, the textual representation of ‘نبات مروج’، ‘*Lawn*’ includes the noise words ‘استعمل’، ‘*use*’, ‘تجاوز’، ‘*exceed*’, ‘ارتفاع’، ‘*height*’, ‘طرق’، ‘*ways*’, ‘امامي’، ‘*frontal*’, and ‘مسافة’، ‘*distance*’. In contrast, the features of the synset include important words that are semantically related to the synset. This meant the ESA model had a robust semantic representation vector for each synset. Therefore, both sources should be used in the mapping, as the semantic knowledge gives more semantic indications than depending on only one source.

8 Extending AWN’s coverage

The major limitation of the current AWN is the low coverage of concepts compared to WordNets for other languages, since most Arabic concepts have not been inserted

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

into the AWN database. This limitation constitutes a big obstruction for using AWN as the source for knowledge-based approaches. In this section, we adapt the mapping method for extending the coverage of AWN with a semi-structured source with high coverage (AWP). Based on the full mapping results, the articles in AWP have been classified as *mapped* or *not-mapped*. The not-mapped articles contain the concepts that are not found in the current AWN database. Additionally, the synsets in PWN can also be divided according to their manual annotations into two groups: *transferred* (synsets that have been linked to Arabic synsets) and *not transferred*. The coverage is extended by adding the Arabic concepts from the not-mapped articles as new synsets in AWN. To achieve this task, the mapping method is applied to the not-mapped articles and the not-transferred synsets to establish links between them. Since the mapping depends on bilingual and monolingual features, the not-mapped articles are filtered to exclude articles that do not have bilingual features.

8.1 Adapting the mapping method

The mapping method is adapted to classify each concept in the not-mapped articles as *inserted* or *not-inserted* (instead of *mapped* or *not-mapped* as in the main method) as follows. For an article, the bilingual features are used to retrieve the set of English synsets in PWN as the candidates using the matching technique. Then, the semantic hierarchy of Arabic is used in parallel with the PWN semantic hierarchy to check whether Arabic features can be extracted for each English synset in the candidate set. For each English synset, Arabic features can be extracted when the parents of this synset have identical Arabic synsets. Consequently, the mapping can be established if all English synsets of an article have at least one parent in the Arabic semantic hierarchy, otherwise the article is ignored in the mapping and extension without further processing. In addition to using Arabic features in measuring semantic similarity, this condition for having parents is also intended to ensure that a potential inserted synset is not isolated in the Arabic hierarchy.

The final output of candidate extraction for an article is the positions in the Arabic semantic hierarchy that align to the positions of candidates in the English hierarchy. To illustrate this step, suppose the given pair [*rectifier*, مقوم التيار] includes the titles of the article in Arabic and English (Figure 10). The English lemma *rectifier* is used only to extract two candidates with their parents (*person* and *electrical device*) in the semantic taxonomy of PWN. Because these two parents have identical nodes (أداة كهربائية and شخص) in the Arabic semantic hierarchy, there are two positions in AWN (Arabic candidates) for the article, as seen in Figure 10. The semantic similarity between the article and each position candidate is computed by integrating the bilingual and only article-based (monolingual) measures. In the article-based measure, the position candidate is represented as the bag-of-words set from the extracted gloss of this position (hierarchal relations) and the non-hierarchal relations of its parents and siblings. Finally, the article is classified based on the similarity scores as *inserted* or *not-inserted*.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

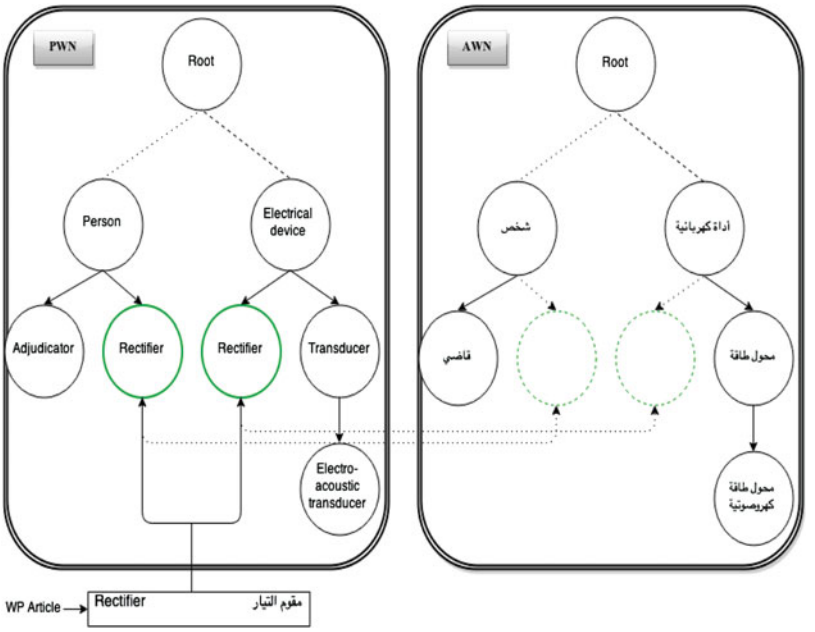


Fig. 10. (Colour online) Example of detecting the positions of an article in AWN.

8.2 Pruning rules

For each article that has been classified as inserted, one node is inserted temporarily in the Arabic semantic hierarchy according to the detected position. To improve the reliability, we introduced two rules to evaluate automatically the inserted nodes based on the hierarchical relations. The rules are applied to the inserted nodes to decide whether the article can be added to AWN as a synset. The first rule depends on the inheritance terming relation between the parent and its child in the hierarchy. The terming relation is when the child concept is labeled by extending the term of its parent concept. For the inserted node, the node of the article has the terming relation if and only if the title of this article is a multiword expression starting with the synonym word of the parent of the inserted position. The following examples are concepts that have the terming relation:

Parent concept:	إلتهاب	inflammation	مرض	disease
Child concept:	التهاب الصلبة	scleritis	مرض مناعي ذاتي	autoimmune disease

AWP can be constructed in a hierarchical way as an acyclic directed graph using the category features if the relation between a category and its subcategories is identical to the hyponymy semantic relation in WordNet. In this graph, the category 'المحتويات' '*Contents*' is the superordinate of all other categories (the root of the hierarchy). The AWP category graph consists of 38,018 nodes and 95,118 edges after removing the cycles using the colored depth-first-search technique (Zesch and Gurevych 2010). Articles are added to this graph as leaf nodes based on the article–category links

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Table 14. *Results of extending AWN's coverage*

Rules	# Concepts	Percentage
Terming rule	1,722	40 per cent
Compatibility rule	2,287	53 per cent
All	4,009	93 per cent

(links from an article to its categories). This graph is considered to be the semantic hierarchy of AWP.

The second rule depends on the compatibility hierarchal relation that can be captured from the semantic hierarchies of AWN and AWP based on the full mapping results. For an article with its target position node, the related synsets are retrieved from the semantic hierarchy of AWN with their corresponding AWP articles that have been linked using the mapping method. Formally, let y be the position node for the article x, the related synsets (RS) of y with the corresponding articles (RA) are defined as the semantic features of the node in the two sources as: $RS(y) = \{y_1, y_2, \ldots, y_m\}andRA(y) = \{x_1, x_2, \ldots, x_m\}$, where x_i is the identical article of the synset y_i on the basis of the full mapping. The related synsets of a node are extracted from the hierarchical relations with respect to the generality and specify notions (as in the extraction of glosses). Based on the semantic hierarchies of AWN and AWP, the relation between x_i and y_i is defined as the distance between the length of the shortest path from x to x_i in the AWP graph and the length of the shortest path from y to y_i in the AWN graph. Therefore, the compatibility between x and y can be determined by the distance vector of the relations among the related synsets with their identical articles as

$$DV(x, y) = |Lx_1 - Ly_1|, |Lx_2 - Ly_2|, \ldots., |Lx_m - Ly_m|$$

where Lx_i/Ly_i is the length of the shortest path between x/y and x_i/y_i in the AWP/AWN semantic graph. Based on this vector, an article is considered compatible to the inserted node if all elements in the distance vector range between 0 and 3.

8.3 Results

In this section, we report the results of extending the coverage of AWN from Wikipedia articles using monolingual and bilingual features. The proposed method has been applied to 47,957 not-mapped articles that have bilingual features in AWP. Based on the semantic hierarchies of AWN and PWN, these articles have been filtered to contain only 4,326 concepts with Arabic parents in the candidate extraction step. The evaluation rules are applied on the articles (4,313) that have been classified as inserted. Table 14 shows the number of entities added to the AWN database after applying the evaluation rules.

9 Conclusion

In this paper, we introduced an automatic method that maps a specific AWN synset to its identical AWP article based on the monolingual and bilingual features.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

For measuring the similarity of Arabic features, we also proposed a new unit-based technique that exploits AWN and AWP as the knowledge sources for modeling the semantic similarity in two directions. The proposed method can be directly applied to cross-lingual WordNets for mapping and enrichment. Several experiments were conducted on the Arabic mapping data set to evaluate the proposed method. For comparison, we adapted five monolingual methods from the literature to Arabic knowledge sources. The comparison shows that the proposed method significantly outperformed the state-of-the-art methods. The proposed method used cross-lingual features and achieved a higher accuracy (93.6 per cent) than the state-of-the-art methods (monolingual-based mapping), whose accuracy ranged between 74.1 per cent and 80.6 per cent. Cross-linguistically, we concluded that the leveraging of bilingual features is useful for improving the mapping for either the matching or the semantic similarity between the synsets and articles. Due to the incompatibility of terming the concepts in Arabic, the improvement ratio from using the bilingual features in candidate extraction was higher than in measuring the similarity.

For measuring the semantic similarity of Arabic features, the unit-based measure, which integrates two measures (synset-based and article-based), was more accurate than its individual measures, as well as the semantic measures in the state-of-the-art mapping methods. In other words, the exploitation of the semantics in AWN and AWP together could provide more semantic evidence than using the semantics from only one of them. By comparing the knowledge sources, the article-based measure, which depends on modeling the semantics of AWP, achieved better evaluation results than the synset-based measure, which depends on AWN as the source of semantics. For the similarity techniques, the unit-based technique was better than the TB, which modeled the semantic similarity between a synset and an article as the similarity between two texts. This strengthens our motivation to use the cross-lingual features and the cross-semantic measures in mapping instead of using only monolingual features and WordNet-based semantics, as in the state-of-the-art methods.

As shown in the full mapping results, the proposed method has been adapted to increase the coverage of AWN by adding new concepts (synsets) from not-mapped articles in Wikipedia. From 4,326 articles with their candidates that have bilingual and monolingual features, the method successfully classified 4,313 as inserted concepts (articles that can be inserted into the Arabic semantic hierarchy). Based on an automatic evaluation of the inserted concepts, 4,009 concepts were added as synsets to the current Arabic nominal synsets in the AWN database.

For future work, we are planning to propose a supervised algorithm to evaluate the inserted entities. The proposed method could be improved to handle articles with isolated candidates. The proposed method also could be used for enriching the existing WordNet synsets (by extracting the glosses and semantic relations) from the semantic information in Wikipedia. The proposed method could be exploited to address the bottleneck of building a sense-tagged corpus for Arabic supervised word-sense disambiguation.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Acknowledgements

This work was funded by the Ministry of Higher Education in Malaysia under the grant no. (FRGS/1/2012/SG0S/UKM/02/14).

References

Abouenour, L., Bouzoubaa, K., and Rosso, P. 2010. Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet. *Workshop on Language Resources and Human Language Technologies for Semitic Languages Status, Updates, and Prospects (LREC-2010) Conference*, Malta, pp. 27–31.

Agirre, E., and Soroa, A. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics. Athens, Greece, pp. 33–41.

Al-Asal, M. S., and Smadi, O. M. 2012. Arabicization and Arabic expanding techniques used in science lectures in two Arab universities. *Perspectives in the Arts and Humanities Asia* **2**(1): 15–38.

Alhanini, Y., and Ab Aziz, M. J. 2011. The enhancement of Arabic stemming by using light stemming and dictionary-based stemming. *Journal of Software Engineering and Applications* **4**(9): 522–26.

Alkhalifa, M., and Rodríguez, H. 2008. Automatically extending named entities coverage of Arabic WordNet using Wikipedia. *International Journal on Information and Communication Technologies* **1**(1): 1–17.

Atserias, J., Climent, S., Rigau, G., and Rodriguez, H. 1997. Combining multiple methods for the automatic construction of multilingual WordNets. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP-1997)*, Association for Computational Linguistics (ACL). Tzigov Chark, pp. 143–49.

Cilibrasi, R. L., and Vitanyi, P. M. B. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* **19**(3): 370–83.

Cucerzan, S. 2007. Large-scale Named Entity disambiguation based on Wikipedia data. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2007)*, Association for Computational Linguistics. Prague, Czech Republic, pp. 708–16.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. 2006. Building a wordnet for arabic. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Citeseer. Genoa - Italy, pp. 29–34.

Fellbaum, C. 1998. *WordNet: an Electrical Lexical Database*. Cambridge, MA: The MIT Press.

Fernando, S., and Stevenson, M. 2012. Mapping WordNet synsets to Wikipedia articles. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, European Language Resources Association (ELRA). Istanbul, Turkey, pp. 590–96.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, Morgan Kaufmann. Hyderabad, India, pp. 1606–11.

Hassan, S., and Mihalcea, R. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. Singapore, pp. 1192–201.

Hassan, S., and Mihalcea, R. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of AAAI 2011 (25th AAAI Conference on Artificial Intelligence)*, Association for the Advancement of Artificial Intelligence. San Francisco, pp. 884–89.

Kashgary, A. D. 2011. The paradox of translating the untranslatable: Equivalence vs. non-equivalence in translating from Arabic into English. *Journal of King Saud University-Languages and Translation* **23**(1): 47–57.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database* **49**(2): 265–83.

Li, Y., Bandar, Z. A., and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering,* **15**(4): 871–82.

Matuschek, M., and Gurevych, I. 2013. Dijkstra-WSA: a graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics* **1**(1): 151–64.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* **67**(9): 716–54.

Mihalcea, R. 2007. Using Wikipedia for automatic word sense disambiguation. *In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics. Rochester, New York, pp. 196–203.

Milne, D., and Witten, I. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy,* Association for the Advancement of Artificial Intelligence. Chicago, USA: AAAI Press, pp. 25–30.

Navigli, R., and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193**(12): 217–50.

Niemann, E., and Gurevych, I. 2011. The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS-2011)*, Citeseer. Oxford, UK, pp. 205–14.

Paul, P. 1978. *Longman Dictionary of Contemporary English.* England: Longman Group Limited.

Pilehvar, M. T., and Navigli, R. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)*, Association for Computational Linguistics. Baltimore, Maryland, pp. 468–78.

Pirró, G., and Euzenat, J. 2010. A feature and information theoretic framework for semantic similarity and relatedness. In *Proceedings of the 9th International Semantic Web Conference (ISWC-2010)*, Springer. Shanghai, China, pp. 615–30.

Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics. Uppsala, Sweden, pp. 1522–31.

Pradet, Q., de Chalendar, G., and Desormeaux, J. B. 2014. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia, pp. 32–39.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, pp. 448–53.

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M. A., Black, W., Elkateb, S., Kirk, J., and Pease, A. 2008. Arabic wordnet: Current state and future extensions. In *Proceedings of the 4th Global WordNet Conference*, Citeseer. Szeged, Hungary, pp. 1–20.

Roget, P. M. 1911. *Roget'S International Thesaurus*, 1st ed. New York, USA: Thomas Y. Crowell Co.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. *Advances in Web Intelligence*, pp. 380–86. Lodz, Poland: Springer.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press

Saif, A., Ab Aziz, M. J., and Omar, N. 2013. Measuring the compositionality of Arabic multiword expressions. *Soft Computing Applications and Intelligent Systems*, pp. 245–56. Shah Alam, Malaysia: Springer.

Sánchez, D., Batet, M., and Isern, D. 2011. Ontology-based information content computation. *Knowledge-Based Systems* **24**(2): 297–303.

Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. *16th European Conference on Artificial Intelligence (ECAI-2004), Including Prestigious Applicants of Intelligent Systems*, IOS Press. Valencia, Spain, pp. 1089–90.

Suchanek, F. M., Kasneci, G., and Weikum, G. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, ACM. Banff, Canada, pp. 697–706.

Toral, A., Munoz, R., and Monachini, M. 2008. Named entity wordnet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Citeseer. Marrakech, Marocco, pp. 741–47.

Vossen, P. 1998. *A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, P., Soroa, A., Zapirain, B., and Rigau, G. 2012. Cross-lingual event-mining using wordnet as a shared knowledge interface. *6th Global WordNet Conference*, Publ. Tribun EU. Matsue, Japan, pp. 382–89.

Wolf, E., and Gurevych, I. 2010. Aligning sense inventories in wikipedia and wordnet. In *Proceedings of the 1st Workshop on Automated Knowledge Base Construction*, Citeseer. Grenoble, France, pp. 24–28.

Wu, Z., and Palmer, M. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics. Las Cruces, New Mexico, pp. 133–38.

Zesch, T., and Gurevych, I. 2010. Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. *Natural Language Engineering* **16**(1): 25–59.

Zhang, Z., Gentile, A. L., and Ciravegna, F. 2012. Recent advances in methods of lexical semantic relatedness–a survey. *Natural Language Engineering* **1**(1): 1–69.

Zhou, Z., Wang, Y., and Gu, J. 2008. A new model of information content for semantic similarity in WordNet. *2nd International Conference on Future Generation Communication and Networking Symposia (FGCNS-2008)*, IEEE. Hainan Island, China, pp. 85–89.

https://doi.org/10.1017/S1351324915000376 Published online by Cambridge University Press