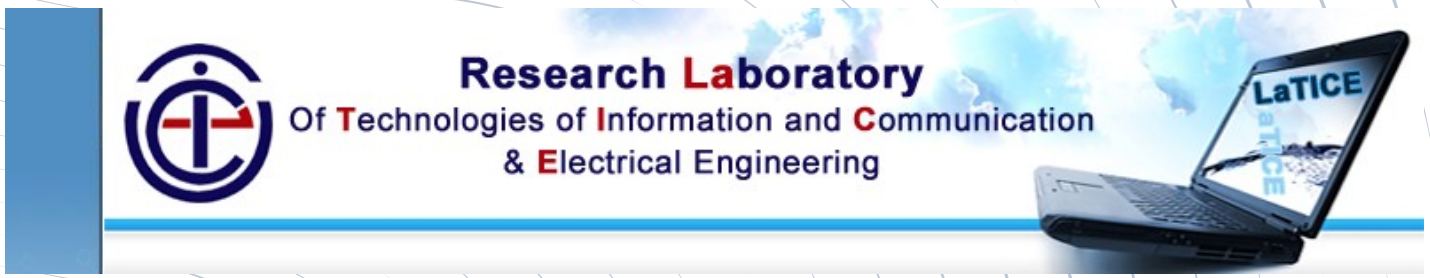
Derivational Relations in Arabic WordNet

December 30, 2017

Mohamed Ali Batita, and Mounir Zrigui
batitamohamedali@gmail.com
mounir.zrigui@fsm.rnu.tn

Research Laboratory of Technologies of Information and Communication & Electrical Engineering
-
unity of Monastir



Plan

Introduction

General introduction

Wordnet is a lexical database built of **synsets**.

Synset is a group of words that share the same concept and the same part-of-speech (POS).

Synsets are interconnected with different relations (Hyponym, antonym, synonym. . .).

The majority of wordnets include the following lexical categories: nouns, verbs, adjectives, and adverbs.

Overview of the AWN

3

Arabic WordNet

AWN was created for Modern Standard Arabic (MSA) language.

Has been created in 2006, and extended in 2015.

Followed the development process of Princeton WordNet and EuroWordNet.

Overview of the AWN

Problems

- Top-down procedure
  1. Translation of the Princeton WordNet's core.
  2. Extension through more specific concepts.

Overview of the AWN

Problems

- Top-down procedure
  1. Translation of the Princeton WordNet's core.
  2. Extension through more specific concepts.
- No cross-POS links.

Overview of the AWN

First version of AWN

- 21,813 words.
- 9,698 synsets.
- 143,715 links.
- 6 types of relations.

Overview of the AWN

Second version of AWN

- 23,841 words.
- 11,269 synsets.
- 161,705 links.
- 22 types of relations (5 inter-languages types.)

Overview of the AWN

LMF version[1]

- 60,157 words.
- 8,550 synsets.
- 41,136 links.
- **4** types of relations.

[1] Extended version of The AWN structured under the LMF standard

Overview of the AWN

Relations in the LMF version

Relation	Frequency
Hyponym / hypernym	19,806
HasInstance / isInstance	549
Similar	412
Antonym	14

Part 1: Wordnets

Work	Wordnet	Authors	Year
Provided a list of affixes, constructed morphosemantically related pairs, and finally linked the synsets via the right relations.	Turkish Czech Croatian	Bilgin *et Al.* Pala *et Al.* Šojat *et Al.*	2004 2007 2014
Added morphological relations between derived pairs of words sharing the same steam.	English Romanian	Fellbaum *et Al.* Mititelu *et Al.*	2007 2012
Automatically transferred some of the relations form PWN and added some language-specific derivational relations.	Bulgarian Serbian	Koeva *et Al.* Koeva *et Al.*	2008 2008
Developed a tool to bootstrap derivational relations from plWordNet.	Polish	Piasecki *et Al.*	2012

Part 2: Others sources

Authors	Year	Approach
Can *et Al.*	2009	Unsupervised method based on different POS to produce morphological rules.
Bernhard	2010	Two unsupervised methods; one to group words into hierarchical families and the other to build a semantic network from these words.
Shaalan	2010	A case study of 3 systems and 4 tools used rule-based approaches and gave satisfactory results.
Tribout *et Al.*	2012	Automatic construction of a morphological resource *(Verbagent)* that groups verbs with their agents.
Zaghouani *et Al.*	2016	Construction of *AMPN*, a lexical resource for the Arabic language, based on morphological patterns.

Our approach
Description

- We adopted an approach based on morphological transformation rules.
- Each rule is based on the *POS* and the *pattern* of the lexical entries.
- We only worked on the words that already exist in the AWN (we did not add new words.)
- The relation happens only when two lexical entries:
  - Belong to AWN.
  - Have the same root.
  - Have a rule that allows the transformation.

Our approach
Description

Majors steps

1. Gather the lexical entries that share the same root (into a *bag of words*).
2. Apply the transformation rules to assign appropriate relations between the words in the same bag.

Clustering words into *bag of words*

First step

This step is based on the root of each word and allows us to:

- Eliminate the underivatized words like named entities.
- Keep the apolistic generic noun (حيوان ، ملعب *ḥywān, mlʿb*).
- Distinguish words that share the same root but no relationship in the stage of meaning (شجرٌ ، شِجار *šğrun, šiğār*).
- Verify the POS of the remaining words.

Transformation rules
Second step

Assigning a relation to a pair of words is based on the change of the *POS*.

- For example, the relation between *write* كتب *ktb* and *writer* كاتب *kātb* is *active participle* فاعل إسم *ʾism fāʿl*.

Transformation rules
Second step

For this matter, we work out **4** sets of rules:

1. **HasDerivedNoun** between *nouns* (كلب، كلاب) *(klb, klāb)* (dog, dogs).
2. **Relatedness** between *noun and adjective* (سياسة، سياسي) *(syāst, syāsy)* (politics, politician).
3. **HasDerivedVerb** between *verbs* (فتح، إفتتح) *(fṯḥ, ʾiftṯḥ)* (open, initiate).
4. **ActiveParticiple**, **PassiveParticiple**, **Location**, **Time** and **Instrument** between *verb and noun*.

Transformation rules
Second step

7

The first, second, and third rules are easy to distinguish

- For the first, and the third, the key is the same POS.
- For the second, the key is the POS switch.

[1] knowing that, a triliteral verb is a verb that only has its root letters.

Transformation rules
Second step

The first, second, and third rules are easy to distinguish

- For the first, and the third, the key is the same POS.
- For the second, the key is the POS switch.

The fourth subset of rules makes some ambiguity

- We classified the verbs into two groups: **triliteral**[1] ثلاثي *ṯlāṯy* and **non-triliteral** غير ثلاثي *ġyr ṯlāṯy*

[1] knowing that, a triliteral verb is a verb that only has its root letters.

Transformation rules

Second step

Relation	Class	Pattern	Example
ActivePariticiple إسم فاعل *ʾism fāʿl*	Triliteral	فاعل *fāʿl* (1A2i3u)	حامد ، حمد *ḥmd, ḥāmd*
		2nd letter is *weak* (فعل أجوف *fʿl ʾağwf*) → ئ *ʾy hamza*	فاح ، فائح *fāḥ, fāʾyḥ*
		3rd letter is *weak* (فعل ناقص *fʿl nāqṣ*) → ي *y ya*	دعا ، دعي *dʿā, dʿy*
	Non-triliteral	مُفعِلْ *mufʿil* (mu1a2i3u)	علم ، معلم *ʿlm, mʿlm*
PassiveParticiple اسم مفعول *āsm mfʿwl*	Triliteral	مفعول *mfʿwl* (ma12u3u)	شرب ، مشروب *šrb, mšrwb*
		م *m* + nom dverbal (مصدر) *(mṣdr)*	قال ، مقول *qāl, mqwl*
	Non-triliteral	مفاعل *mfāʿl* (m1A2i3u)	بارك ، مبارك *bārk, mbārk*
Location اسم مكان *āsm mkān*	Triliteral	مفعَل *mfʿal* (ma12a3u)	طبخ ، مطبخ *ṭbḫ, mṭbḫ*
Time اسم زمان *āsm zmān*	Triliteral	مفعِل *mfʿil* (ma12i3u)	غرب ، مغرب *ġrb, mġrb*
Instrument اسم ألة *āsm ʾalh*	-	مفعْل *mfʿl* (mi12a3u)	عول ، معول *ʿwl, mʿwl*
		مفعلة *mfʿlh* (mi12a3h)	قلم ، مقلمة *qlm, mqlmh*
		مفعال *mfʿāl* (mi12A3u)	فتح ، مفتاح *fṭḥ, mftāḥ*
		فعالة *fʿālh* (1i2A3h)	غسل ، غسالة *ġsl, ġsālh*

Transformation rules

Second step

Relation	Class	Pattern	Example
ActiveParticiple إسم فاعل *ʾism fāʿl*	Triliteral	فاعل *fāʿl* (1A2i3u)	حامد ، حمد *ḥmd, ḥāmd*
		2nd letter is *weak* (فعل أجوف *fʿl ʾağwf*) → ئ *ʾy hamza*	فاح ، فائح *fāḥ, fāʾyḥ*
		3rd letter is *weak* (فعل ناقص *fʿl nāqṣ*) → ي *y ya*	دعا ، دعي *dʿā, dʿy*
	Non-triliteral	مُفعِلْ *mufʿil* (mu1a2i3u)	علم ، معلم *ʿlm, mʿlm*
PassiveParticiple اسم مفعول *āsm mfʿwl*	Triliteral	مفعول *mfʿwl* (ma12u3u)	شرب ، مشروب *šrb, mšrwb*
		م *m* + nom dverbal (مصدر) *(mṣdr)*	قال ، مقول *qāl, mqwl*
	Non-triliteral	مفاعل *mfāʿl* (m1A2i3u)	بارك ، مبارك *bārk, mbārk*
Location اسم مكان *āsm mkān*	Triliteral	مفعَل *mfʿal* (ma12a3u)	طبخ ، مطبخ *ṭbḫ, mṭbḫ*
Time اسم زمان *āsm zmān*	Triliteral	مفعِل *mfʿil* (ma12i3u)	غرب ، مغرب *ġrb, mġrb*
Instrument اسم ألة *āsm ʾalh*	-	مفعْل *mfʿl* (mi12a3u)	عول ، معول *ʿwl, mʿwl*
		مفعلة *mfʿlh* (mi12a3h)	قلم ، مقلمة *qlm, mqlmh*
		مفعال *mfʿāl* (mi12A3u)	فتح ، مفتاح *fṭḥ, mftāḥ*
		فعالة *fʿālh* (1i2A3h)	غسل ، غسالة *ġsl, ġsālh*

Transformation rules

Second step

Relation	Class	Pattern	Example
ActiveParticiple إسم فاعل *ʾism fāʿl*	Triliteral	فاعل *fāʿl* (1A2i3u)	حمد ، حامد *ḥmd, ḥāmd*
		2nd letter is *weak* (فعل أجوف *fʿl ʾağwf*) → ئ *ʾy hamza*	فاح ، فائح *fāḥ, fāʾyḥ*
		3rd letter is *weak* (فعل ناقص *fʿl nāqṣ*) → ي *y ya*	دعا ، دعي *dʿā, dʿy*
	Non-triliteral	مُفعِلْ *mufʿil* (mu1a2i3u)	علم ، معلم *ʿlm, mʿlm*
PassiveParticiple اسم مفعول *āsm mfʿwl*	Triliteral	مفعول *mfʿwl* (ma12u3u)	شرب ، مشروب *šrb, mšrwb*
		م *m* + nom dverbal (مصدر) *(mṣdr)*	قال ، مقول *qāl, mqwl*
	Non-triliteral	مفاعل *mfāʿl* (m1A2i3u)	بارك ، مبارك *bārk, mbārk*
Location اسم مكان *āsm mkān*	Triliteral	مفعَل *mfʿal* (ma12a3u)	طبخ ، مطبخ *ṭbḫ, mṭbḫ*
Time اسم زمان *āsm zmān*	Triliteral	مفعِل *mfʿil* (ma12i3u)	غرب ، مغرب *ġrb, mġrb*
Instrument اسم ألة *āsm ʾalh*	-	مفعْل *mfʿl* (mi12a3u)	عول ، معول *ʿwl, mʿwl*
		مفعلة *mfʿlh* (mi12a3h)	قلم ، مقلمة *qlm, mqlmh*
		مفعال *mfʿāl* (mi12A3u)	فتح ، مفتاح *fṭḥ, mftāḥ*
		فعالة *fʿālh* (1i2A3h)	غسل ، غسالة *ġsl, ġsālh*

Transformation rules

Second step

Relation	Class	Pattern	Example
ActiveParticiple إسم فاعل *ʾism fāʿl*	Triliteral	فاعل *fāʿl* (1A2i3u)	حامد ، حمد *ḥmd, ḥāmd*
		2nd letter is *weak* (فعل أجوف *fʿl ʾağwf*) → ئ *ʾy hamza*	فائح ، فاح *fāḥ, fāʾyḥ*
		3rd letter is *weak* (فعل ناقص *fʿl nāqṣ*) → ي *y ya*	دعي ، دعا *dʿā, dʿy*
	Non-triliteral	مُفعِلْ *mufʿil* (mu1a2i3u)	معلم ، علم *ʿlm, mʿlm*
PassiveParticiple اسم مفعول *āsm mfʿwl*	Triliteral	مفعول *mfʿwl* (ma12u3u)	مشروب ، شرب *šrb, mšrwb*
		م *m* + nom dverbal (مصدر) *(mṣdr)*	مقول ، قال *qāl, mqwl*
	Non-triliteral	مفاعل *mfāʿl* (m1A2i3u)	مبارك ، بارك *bārk, mbārk*
Location اسم مكان *āsm mkān*	Triliteral	مفعَل *mfʿal* (ma12a3u)	مطبخ ، طبخ *ṭbḫ, mṭbḫ*
Time اسم زمان *āsm zmān*	Triliteral	مفعِل *mfʿil* (ma12i3u)	مغرب ، غرب *ġrb, mġrb*
Instrument اسم ألة *āsm ʾalh*	-	مفعْل *mfʿl* (mi12a3u)	معول ، عول *ʿwl, mʿwl*
		مفعلة *mfʿlh* (mi12a3h)	مقلمة ، قلم *qlm, mqlmh*
		مفعال *mfʿāl* (mi12A3u)	مفتاح ، فتح *fṭḥ, mftāḥ*
		فعالة *fʿālh* (1i2A3h)	غسالة ، غسل *ġsl, ġsālh*

Transformation rules
Second step

7

Issues have risen:

- The pattern مفعل *mfʕl* is presented with **4 links** (activeParticiple, location, time, and instrument.)

Transformation rules
Second step

Solution

- With *activeParticiple* link, recognition is easy.
- With others, only from the context.

Validation
Final step

8

validation

- Manually validation by a lexicographer.

Results and assessment

- We implemented this method using Java.
- We had some issues in the verb roots.
- Multiword expressions have to be eliminated.
- Many links associated to nouns (...المثنى، الجموع *ālmṯnā,* *ālğmwʿ...*) are gathered in only one link (*HasDerivedNoun*) to avoid the overlap on them.
- the manually validation took so much time.

Results and assessment

Frequency of new relations

Relation	Frequency
HasDerivedVerb	2,005
ActiveParticiple	1,347
PassiveParticiple	1,004
Location	985
Time	752
Instrument	184
HasDerivedNoun	1,784
Relatedness	804
Total	**8,865**

Conclusion

- Arabic WordNet needs more attention.
- This work is focused on the derivational relations.
- Next step will be the *automatic* validation of all the links in AWN.

Thank you for your attention.