Leveraging Arabic Morphology and Syntax for Achieving Better Keyphrase Extraction

Muhammad Helmy, Dario De Nart, Dante Degl’Innocenti and Carlo Tasso

Artificial Intelligence Laboratory @ University of Udine - http://ailab.uniud.it

Keyphrase Extraction (KPE)

- A Keyphrase (KP) is commonly defined as a short phrase typically consisting in one to three words representing an entity or a concept that is somehow representative of the content of a given text.

- KPE consists in generating a pool of candidate KPs (CKPs) and then select the most relevant ones according to a set of features

- Arabic digital content in the recent years has grown considerably on the Web pushed by the increasing access of Arabic countries to the internet and social media

- Despite the great linguistic differences between Arabic and western languages such as English, most Arabic KPE systems rely on approaches designed for western languages, thus ignoring its rich morphology and syntax

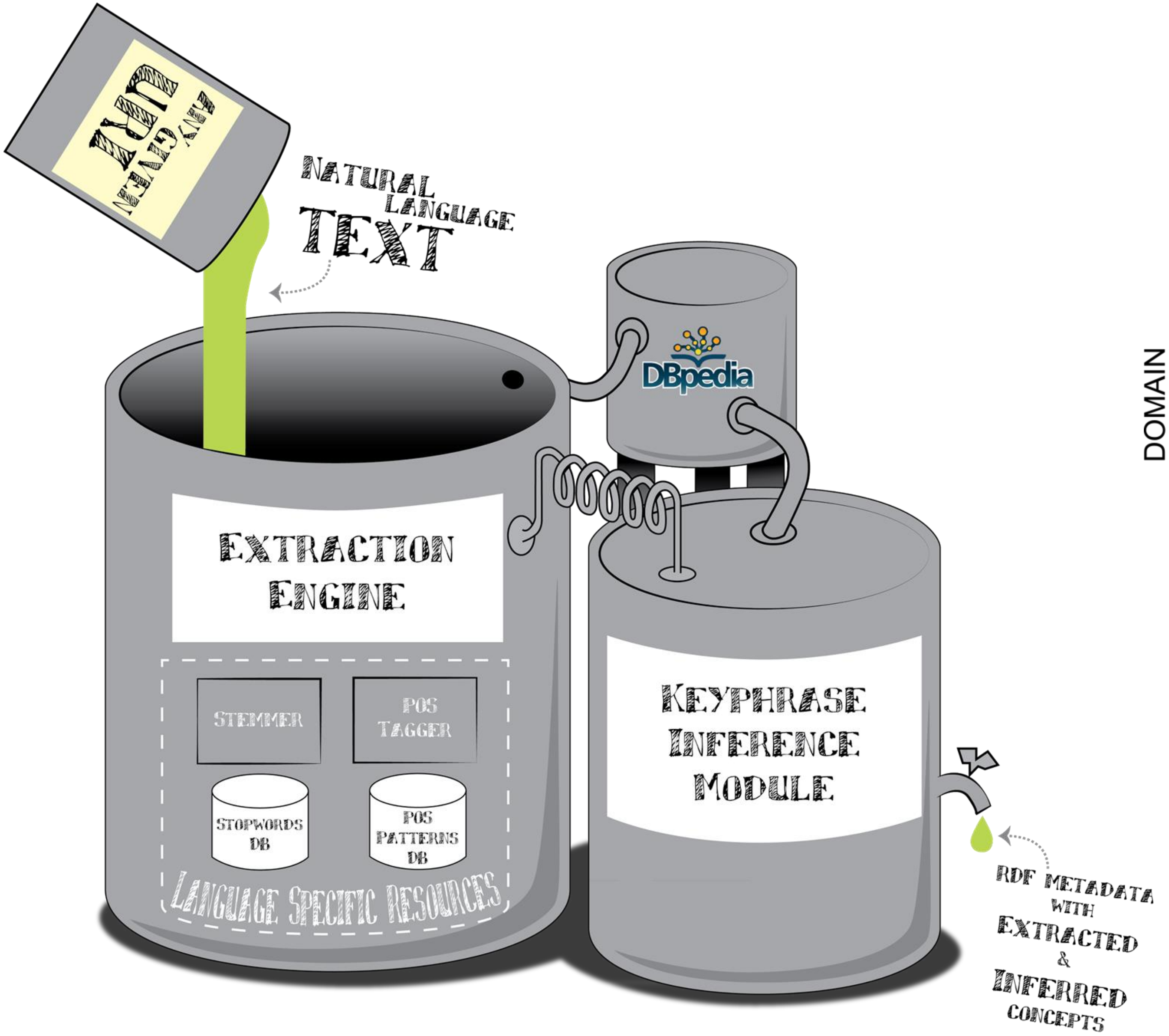
Objectives of the Proposed Approach

- The claim of this work is twofold:

  - Firstly, we believe that approaches tailored on the key characteristics of non-western languages could provide better results than just tuning existing systems originally designed for western languages

  - Secondly, we also believe that a more accurate CKP generation phase, avoiding generation of clearly non-relevant phrases, coupled with a relatively simple selection phase could provide better results than a complex candidate selection relying on a wide array of features coupled with a naïve CKP generation phase

Distiller

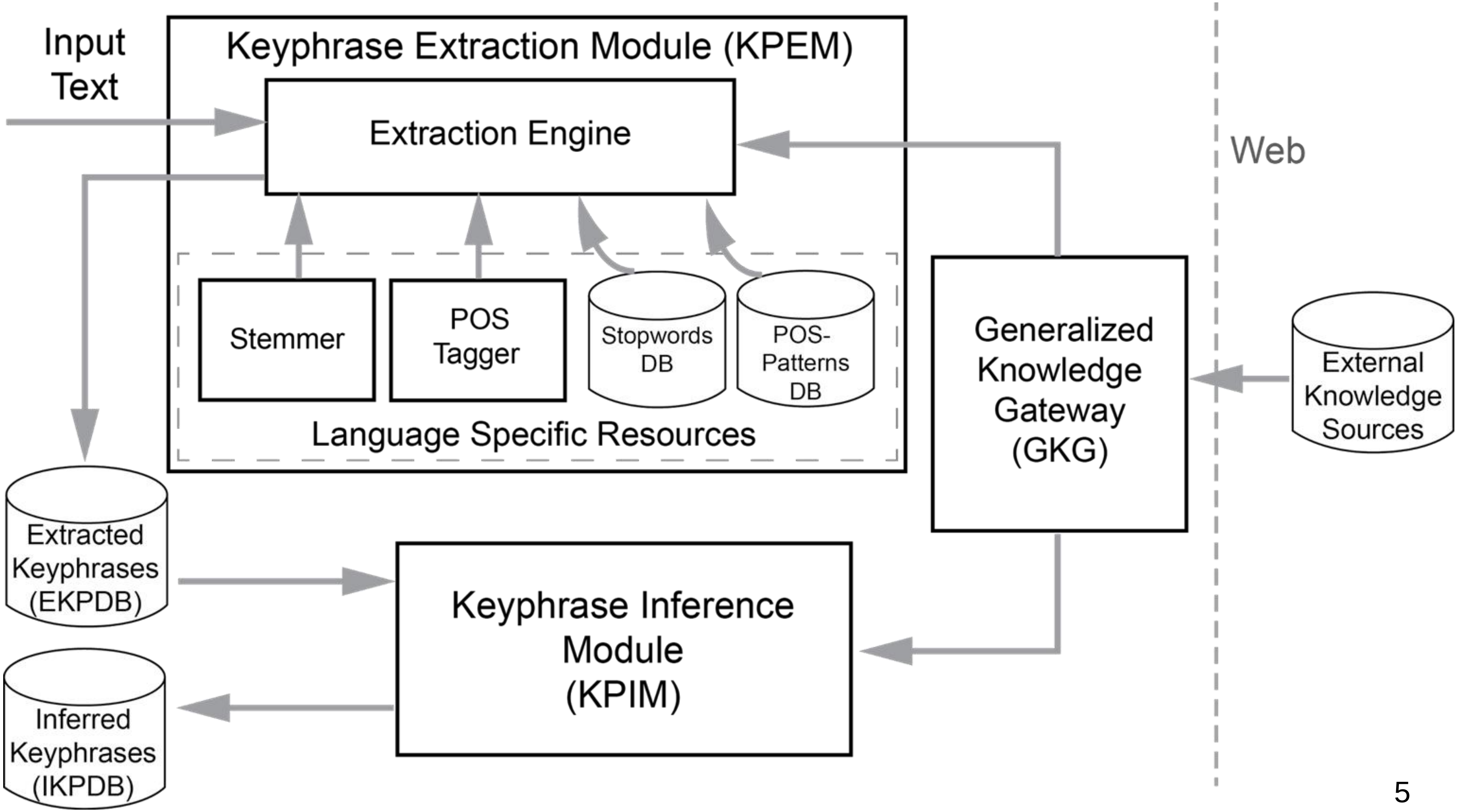


		LANGUAGE	
		independent	dependent
DOMAIN	independent	Statistical Knowledge	Linguistic Knowledge
	dependent	Meta/ Structural Knowledge	Semantic/ Social Knowledge

Works with:

- English
- Italian
- Arabic

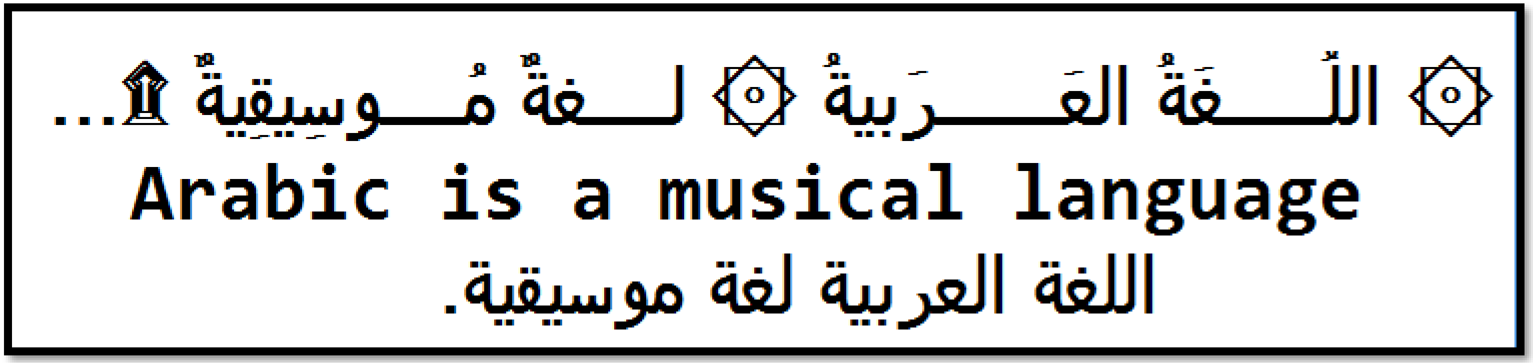
Distiller Architecture



Input Text
Keyphrase Extraction Module (KPEM)
Extraction Engine
Stemmer
POS Tagger
Stopwords DB
POS-Patterns DB
Language Specific Resources
Web
Generalized Knowledge Gateway (GKG)
External Knowledge Sources
Extracted Keyphrases (EKPDB)
Keyphrase Inference Module (KPIM)
Inferred Keyphrases (IKPDB)

Text Cleaning and Normalization

- Cleaning Process is an important step, since it removes the unnecessary characters and symbols from the text and preserves those characters forming the real words of the document

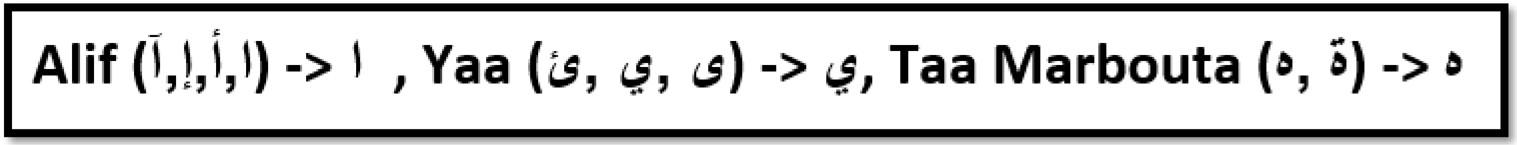


۞ اللُّـــــغَةُ العَـــــرَبيةُ ۞ لـــغةٌ مُـــوسِيقِيةٌ 🕌...

Arabic is a musical language

اللغة العربية لغة موسيقية.

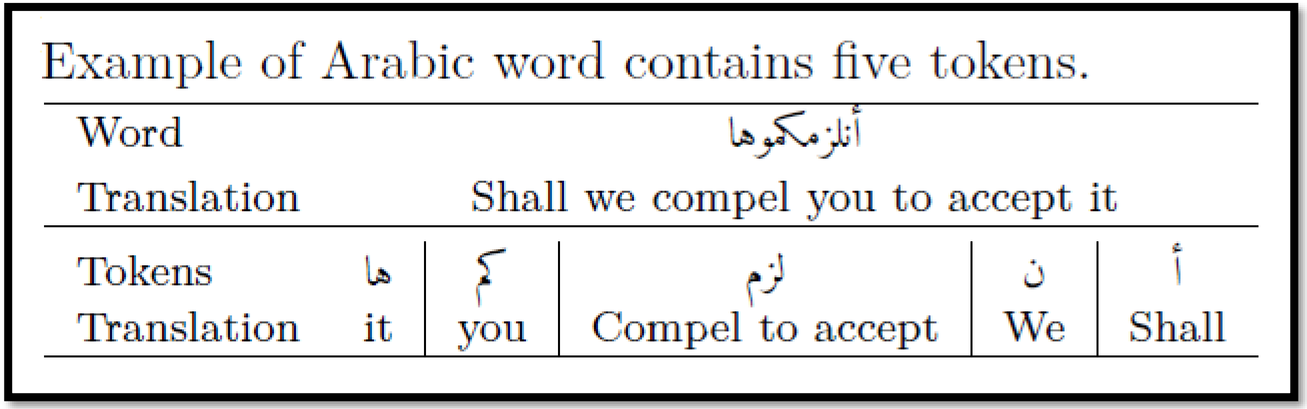
- Normalization unifies the different forms of Arabic letters into a single one throughout the document



Alif (آ,إ,أ,ا) -> ا , Yaa (ئ, ي, ى) -> ي, Taa Marbouta (ه, ة) -> ه

Text Splitting and Segmentation

- Dividing Arabic text into sentences and tokens is not an easy task
- Punctuation marks and whitespaces do not define the boundaries of sentences and words precisely like English
- Moreover, using punctuation marks in MSA is optional and they are rarely used in a strict manner
- Additionally, a single word can hold a complete sentence or a set of concatenated tokens

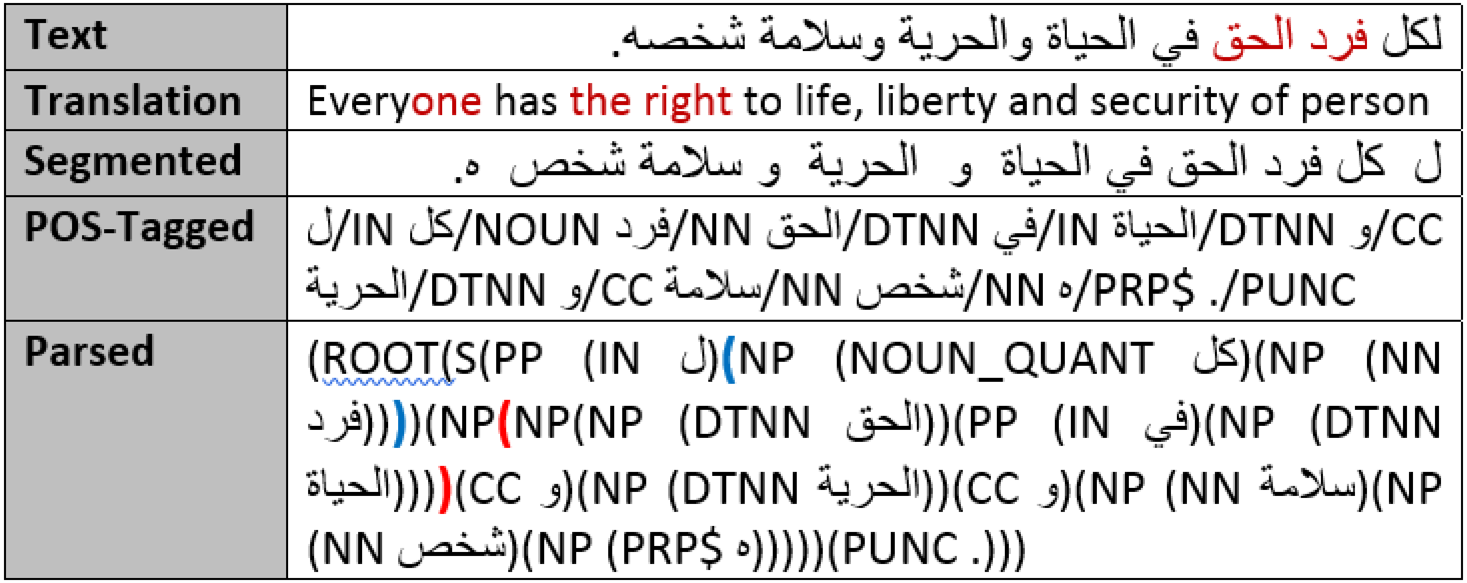


Example of Arabic word contains five tokens.

Word	أنلزمكموها				
Translation	Shall we compel you to accept it				
Tokens	ها	كم	لزم	ن	أ
Translation	it	you	Compel to accept	We	Shall

Text POS-tagging and Parsing

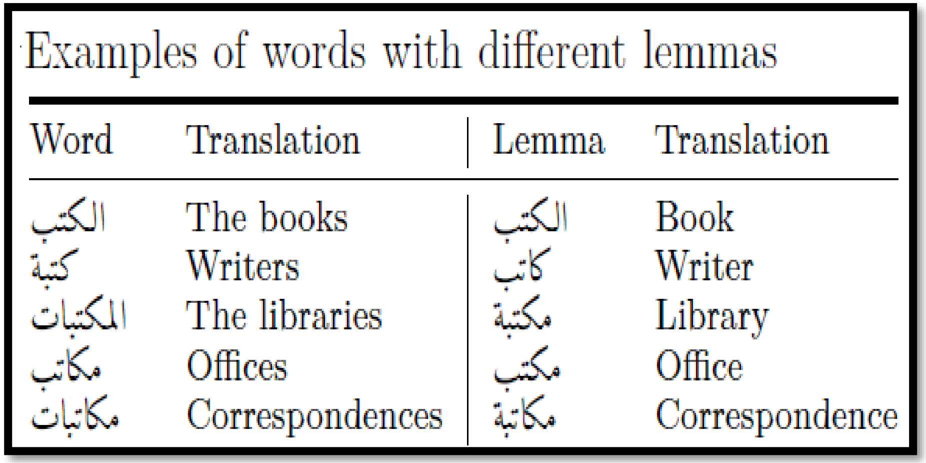
- Every token in the segmented text was assigned a POS-tag depending on its location and context
- POS-tagged text is used to determine CKPs and sentences boundaries
- After that, the text is parsed to detect and generate a list of all NPs



Text	لكل فرد الحق في الحياة والحرية وسلامة شخصه.
Translation	Everyone has the right to life, liberty and security of person
Segmented	ل كل فرد الحق في الحياة و الحرية و سلامة شخص ه.
POS-Tagged	ل/IN كل/NOUN فرد/NN الحق/DTNN في/IN الحياة/DTNN و/CC الحرية/DTNN و/CC سلامة/NN شخص/NN ه/PRP$ ./PUNC
Parsed	(ROOT(S(PP (IN ل)(NP (NOUN_QUANT كل)(NP (NN فرد)))(NP(NP(NP (DTNN الحق))(PP (IN في)(NP (DTNN الحياة))))(CC و)(NP (DTNN الحرية))(CC و)(NP (NN سلامة)(NP (NN شخص)(NP (PRP$ ه)))))(PUNC .)))

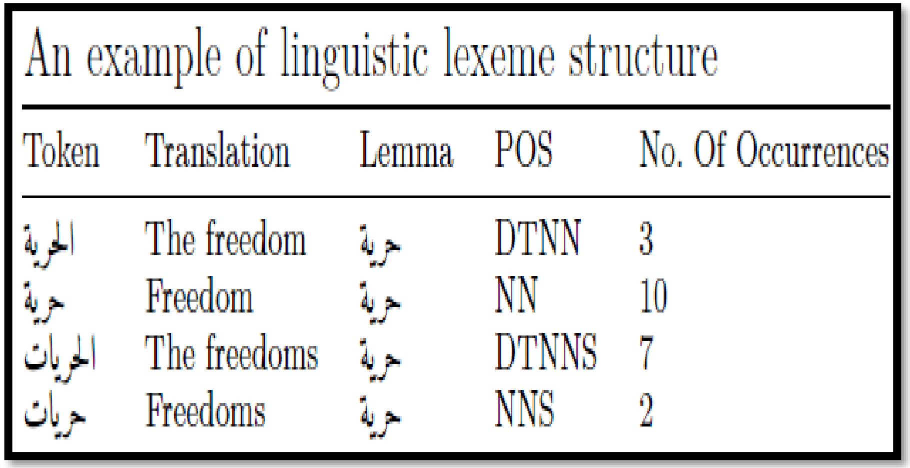
Lemmatization and Tokens Grouping

- Related tokens of the text should be grouped using their basic linguistic form (LF)
- All of the Arabic KPE systems are based on stemming like western languages
- stemming over-reduce the words, so we used lemmatization
- After lemmatizing the text, a list of Linguistic Lexemes (LLs) is generated. Every entry in LLs consists of a set of atomic tokens with the same lemma.



Examples of words with different lemmas

Word	Translation	Lemma	Translation
الكتب	The books	الكتب	Book
كتبة	Writers	كاتب	Writer
المكتبات	The libraries	مكتبة	Library
مكاتب	Offices	مكتب	Office
مكاتبات	Correspondences	مكاتبة	Correspondence



An example of linguistic lexeme structure

Token	Translation	Lemma	POS	No. Of Occurrences
الحرية	The freedom	حرية	DTNN	3
حرية	Freedom	حرية	NN	10
الحريات	The freedoms	حرية	DTNNS	7
حريات	Freedoms	حرية	NNS	2

Generating and Ranking CKPs

- The regular expression:
  (NOUN|ADJECTIVE)((CONNECTOR)?(NOUN|ADJECTIVE)){1,n-1}
  is used to search the POS-tagged text to find the n-gram CKPs with length n

- CKPs are scored according to the following simple score equation (SC):

$$SC(CKP) = \begin{cases} \dfrac{\#LF\_Occ(CKP)}{\#DocTerms} & Length(CKP) = 1 \\ \dfrac{\#LF\_Occ(CKP)+\sum_{k=1}^{n} SC(LGram_k)}{Length(CKP)} & Length(CKP) > 1 \end{cases}$$

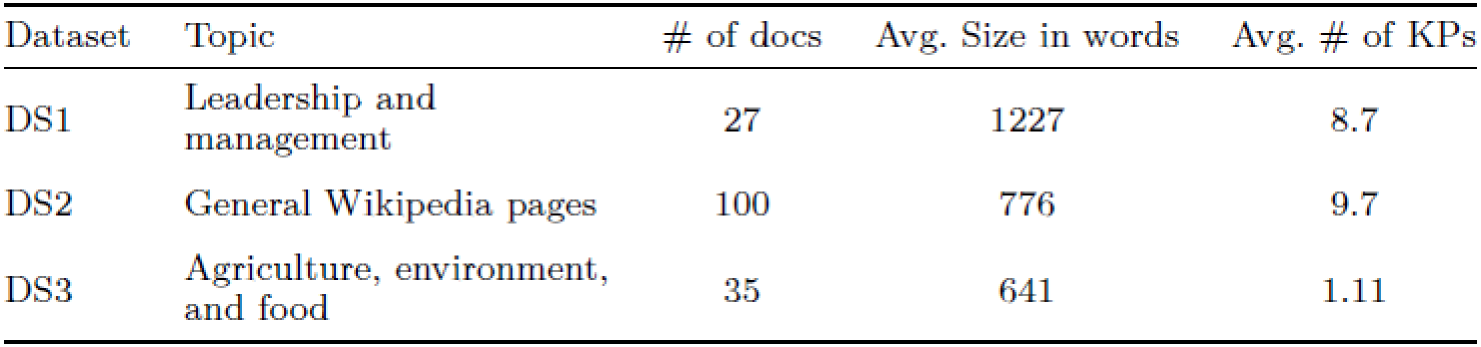
- When the system detects a CKP, it checks whether the text of CKP forms a single NP in the NPs list or not

Evaluation

Dataset

- All of the existing Arabic KPE approaches have been tested and evaluated against datasets built by their authors

- We decided to not build a custom dataset to avoid bias. Instead, we used three datasets already known in the literature

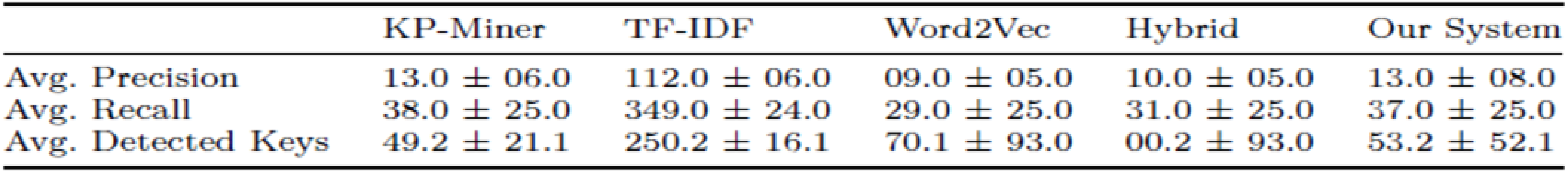


Dataset	Topic	# of docs	Avg. Size in words	Avg. # of KPs
DS1	Leadership and management	27	1227	8.7
DS2	General Wikipedia pages	100	776	9.7
DS3	Agriculture, environment, and food	35	641	1.11

Evaluation

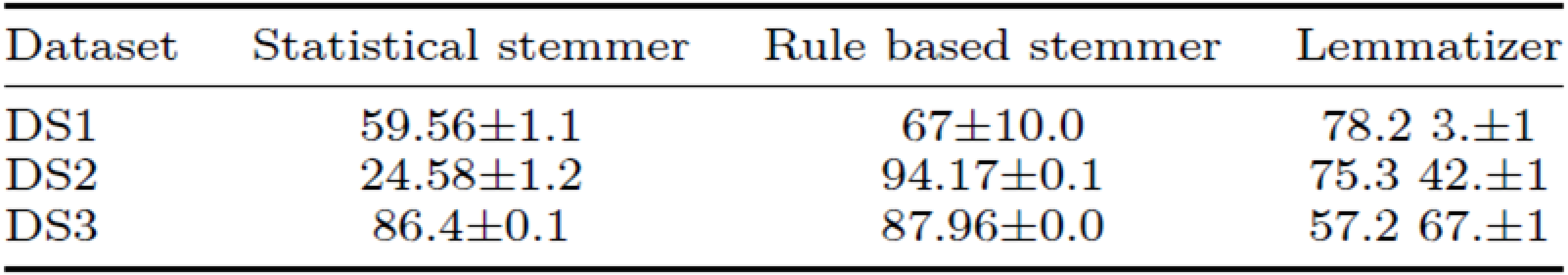
Experimental Results

Comparison between the proposed system and other approaches



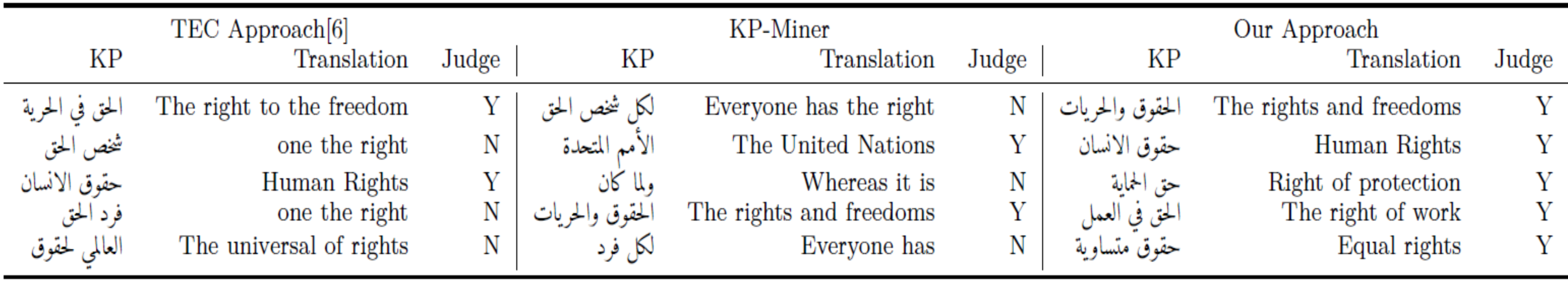
	KP-Miner	TF-IDF	Word2Vec	Hybrid	Our System
Avg. Precision	13.0 ± 06.0	112.0 ± 06.0	09.0 ± 05.0	10.0 ± 05.0	13.0 ± 08.0
Avg. Recall	38.0 ± 25.0	349.0 ± 24.0	29.0 ± 25.0	31.0 ± 25.0	37.0 ± 25.0
Avg. Detected Keys	49.2 ± 21.1	250.2 ± 16.1	70.1 ± 93.0	00.2 ± 93.0	53.2 ± 52.1

Comparison between Arabic-KEA using stemmers and our approach with lemmatizer



Dataset	Statistical stemmer	Rule based stemmer	Lemmatizer
DS1	59.56±1.1	67±10.0	78.2 3.±1
DS2	24.58±1.2	94.17±0.1	75.3 42.±1
DS3	86.4±0.1	87.96±0.0	57.2 67.±1

A comparison for the top-5 KPs extracted by TEC and KP-Miner against the proposed approach



TEC Approach[6]			KP-Miner			Our Approach		
KP	Translation	Judge	KP	Translation	Judge	KP	Translation	Judge
الحق في الحرية	The right to the freedom	Y	لكل شخص الحق	Everyone has the right	N	الحقوق والحريات	The rights and freedoms	Y
شخص الحق	one the right	N	الأمم المتحدة	The United Nations	Y	حقوق الانسان	Human Rights	Y
حقوق الانسان	Human Rights	Y	ولما كان	Whereas it is	N	حق الحماية	Right of protection	Y
فرد الحق	one the right	N	الحقوق والحريات	The rights and freedoms	Y	الحق في العمل	The right of work	Y
العالمي لحقوق	The universal of rights	N	لكل فرد	Everyone has	N	حقوق متساوية	Equal rights	Y

Conclusion and Future Work

- All of the existing Arabic KPE approaches have been tested and evaluated against datasets built by their authors
- Experimental results support our claims, providing evidence that an approach specifically built for Arabic, leveraging linguistic knowledge can outperform western languages based approaches
- Moreover results suggest also that moving the focus from candidate selection to candidate generation could provide a significant performance lift
- Our future work will be focused on coupling our CKP generation approach with a more advanced selection phase and on addressing the problem of linguistic resources shortage

شكراً

Thank you