Discovering Potential Terminological Relationships from Twitter's Timed Content

Mohammad Daoud
Department of Computer Science
American University of Madaba
Madaba, Jordan
m.daoud@aum.edu.jo

Daoud Daoud
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
d.daoud@psut.edu.jo

Abstract

This paper presents a method to discover possible terminological relationships from tweets. We match the histories of terms (frequency patterns). Similar history indicates a possible relationship between terms. For example, if two terms (t1, t2) appeared frequently in Twitter at particular days, and there is a 'similarity' in the frequencies over a period of time, then t1 and t2 can be related. Maintaining standard terminological repository with updated relationships can be difficult; especially in a dynamic domain such as social media where thousands of new terms (neology) are coined every day. So we propose to construct a raw repository of lexical units with unconfirmed relationships. We have experimented our method on time-sensitive Arabic terms used by the online Arabic community of Twitter. We draw relationships between these terms by matching their similar frequency patterns (timelines). We use dynamic time warping as a similarity measure. For evaluation, we have selected 630 possible terms (we call them preterms) and we matched the similarity of these terms over a period of 30 days. Around 270 correct relationships were discovered with a precision of 0.61. These relationships were extracted without considering the textual context of the term.

1 Introduction

Internet users are producing 10,000 Microposts on average every second (internetlivestats 2015). Microposts are short messages containing few sentences written in several languages. These messages tend to talk about time sensitive topics (Grinev, Grineva et al. 2011) (Kwak, Lee et al. 2010). Microposts are rich with terminology (Uherčík, Šimko et al. 2013), not only old and well defined terminology but also newly coined terms (Becker, Naaman et al. 2011).

Building and maintaining an up-to-date terminological repository is very important for several applications (Daoud, Boitet et al. 2010), like machine translation (Vasconcellos, Avey et al. 2001), information retrieval (Peñas, Verdejo et al. 2001)… However, finding terminology (terms and relationships) is a very difficult task (Cabre and Sager 1999), especially for poorly equipped languages, and when the domain is active and changing everyday (new concepts appear every day). Classical approaches in building terminology depend heavily on terminologists and subject-matter experts (Hartley and Paris 1997, Kim, Yang et al. 2005). This approach is very expensive (Gaussier and Langé 1997, Davidson 1998), and it achieves poor coverage (Daoud 2010) because terminologists have limited capability and subject matter experts are rare for contemporary domains. Statistical approaches on the other hand are less expensive, but they need large and processed corpus/corpora. Besides, statistical methods might find a list of candidate terms without relationships, so mapping these terms into a lexical network can be difficult. Microblogs are massive and can solve the problem of the availability of a large textual corpus, however, these microblogs have little textual context (A micropost in Twitter is 140 characters only) and they are usually poorly written (Cornolti, Ferragina et al. 2013).

Proceedings of the Workshop on Cognitive Aspects of the Lexicon,
pages 134–144, Osaka, Japan, December 11-17 2016.

We are working on analyzing terms that appear on microblogs over a period of time to monitor their evolutions. Our idea is that terms with similar histories (frequency patterns over a period of time) are probably similar. For example, if two terms are peaking at the same dates then there is a chance that these terms are used by the internet users synonymously. That way rather than using textual context (which is almost nonexistent in microblogs), we are using historical context to relate between terms. And that will make social media a legitimate source of terminology (terms and relationships). Building a terminological database is still challenging (Roche, Calberg-Challot et al. 2009), because terminology must be standardized and must have a formal body to approve it. We are proposing to extract unconfirmed terminological relationships (preterminology relationships) (Daoud, Boitet et al. 2009, Daoud, Boitet et al. 2009, Daoud, Kageura et al. 2010) rather than standard terminology. Preterminology is considered as raw material for terminology that can be refined to produce standard terminology.

Matching timelines for terms is a classical time series problem, where time series are searched for similarities. There are several approaches to search time series. The performance of these approaches depends on the application (Agrawal, Faloutsos et al. 1993). We use an algorithm originally used for speech recognition called Dynamic Time Warping algorithm (Sakoe and Chiba 1978) with a normalized Euclidean distance function. This approach will not only measure the distance between timelines, but it will consider the slight shifts in the timelines. And this is very suitable for our application because related terms might not peak on the exact same days.

This article is organized as follows; the following section introduces terminology evolution in big data. The third section presents our approach in finding historical similarity between terms. The fourth section shows our data collection method. The fifth section shows the experimental results and evaluation, and finally we will draw some conclusions.

2 Terminology and Preterminology in Big Data

A term is a sign to describe a thought in a particular domain (Sager 1990); this sign is a lexical unit that corresponds to one or more words (Kageura 2002). According to the extended semantic triangle (Suonuuti 1997), a term corresponds to a concept and must have a definition. A terminology is the vocabulary (set of signs) of a domain. Building a term base involves finding precise definitions for each term and connecting terms with relationships. Such process is difficult to achieve in dynamic domains and mediums (Gaussier and Langé 1997, Davidson 1998, Roche, Calberg-Challot et al. 2009). Therefore, we propose to collect preterminology rather that terminology (Daoud 2010). Preterminology is considered as raw material for terminology that can be refined to produce standard terminology. Preterminology incorporates neology (Cabré and Nazar 2011) of new concepts with no standard terms.

Social media posters associate a new concept with a sign (preterm) (Giannakidou, Vakali et al. 2014). This association was not approved by a standardization body and this preterm may not have a specific definition. That is why we call it a preterm rather than a term. A preterm can be processed to produce a term. Social media content may associate two terms (preterms), which can lead to an actual terminological relationship. That is why in this paper we are investigating possible terms (preterms) and their relationships (preterminological relationships). Preterminology can be convenient for useful application such as IR and opinion mining, moreover, it can be used to produce actual terminology.

Extracting knowledge from big data, such as social media generated content, is attracting more and more researchers (Chen, Chiang et al. 2012). Data provided by internet users can be used to find new trends, prevent diseases (Yang, Horneffer et al. 2013), detect crimes (Kandias, Stavrou et al. 2013), and predict future events(Bothos, Apostolou et al. 2010). Extracting terminology or other lexical semantic information from Twitter (Twitter 2015) or social media in general is an ambitious task (Federmann, Gromann et al. 2012). Many succeeded in extracting trending lexical units, finding collocations, classifying tweets, and analyzing positivity/negativity of terms and tweets (Speriosu, Sudan et al. 2011, Zhao, Jiang et al. 2011, Daoud, Alkouz et al. 2015). These attempts consider the textual context of lexical units. However, there is a limitation in using Twitter's textual context as natural language processing of tweets is difficult, especially for Arabic. Therefore, while there is a need and a possibility to extract real-time terminology from tweets, attempts are faced with challenges.. We are proposing a method that considers the textual and the historical context to extract terminological information and relationships.

Traditional terminology has a specific definition that disallows the integration of unconventional resources. That is why a classical standard terminological repository suffers from a lack of linguistic and informational coverage (Gallego Hernández and Herrero Díaz 2014), and it cannot deal flexibly with hidden or absent terminology (Daoud 2010). We suggest extracting unconfirmed terminological relationship between terms (preterms). These possible relationships will have a similarity weight indicating a possible relationship (translation, synonymy, acronym, hyponymy, antonymy, or other).

3 Timeline Similarity

We monitor the frequencies of possible terms each day. We create a timeline for each one. The timeline shows the daily frequencies of the preterm. These timelines illustrate the peaks, bottoms, and possibly the coining date of a preterm. Figure (1) shows the timeline for "اقتحام لاقصى" (Al-Aqsa raid). We can see that the term has peaked on 13 September 2015 with 11,800 frequencies.

The tool used to produce the figure is an online Arabic social media monitoring platform built by the second author. We studied a small set of Arabic preterms and we observed similarities between the timelines of related ones. Figure (2) shows the timelines of "اسعار النفط , اوبك" (OPEC, oil prices). We can see similarity in the frequencies during the period from 25 August 2015 to 29 September 2015. The similarity between terms can occur due to one of the following reasons:

1. Term collocation: terms that co-occur to convey certain meaning, figure 3 shows an example.
2. Event co-occurrence: separate events happened at the same time. Each event has related terms that might produce similar timelines.
3. Same event with different concepts (related terms); Figure (4) shows an example.
4. Same or similar concept with different lexical units (translation, synonymy, acronym, hyponymy, antinomy, hypernymy). Figure (5) shows an example.

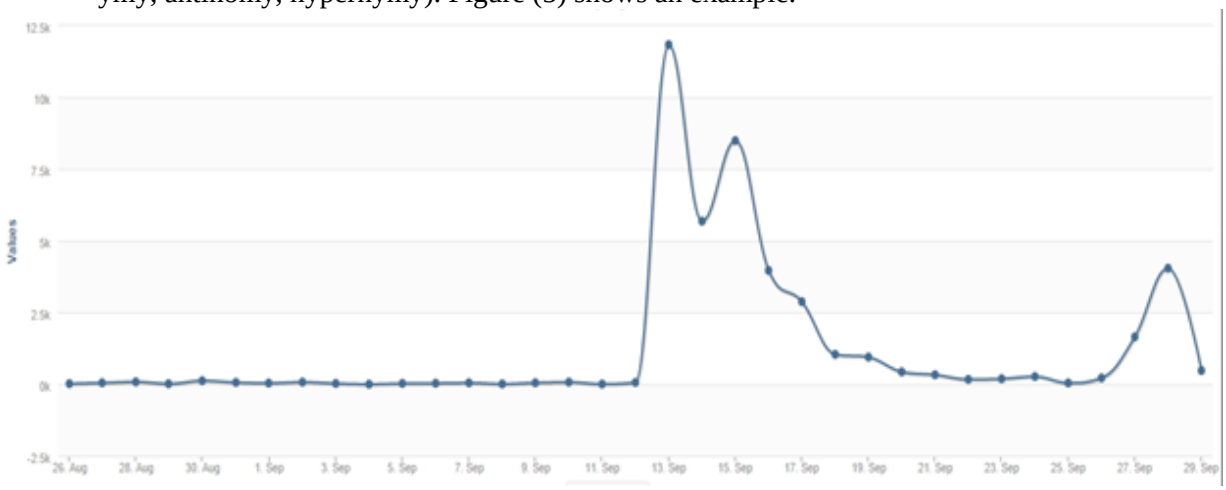

Figure 1. Timeline example for the term "اقتحام لاقصى" (Al-Aqsa raid)

Our objective based on these observations is to search for similar timelines to build a candidate set of relationships between new terms (preterms) extracted from the community of Arabic social media.

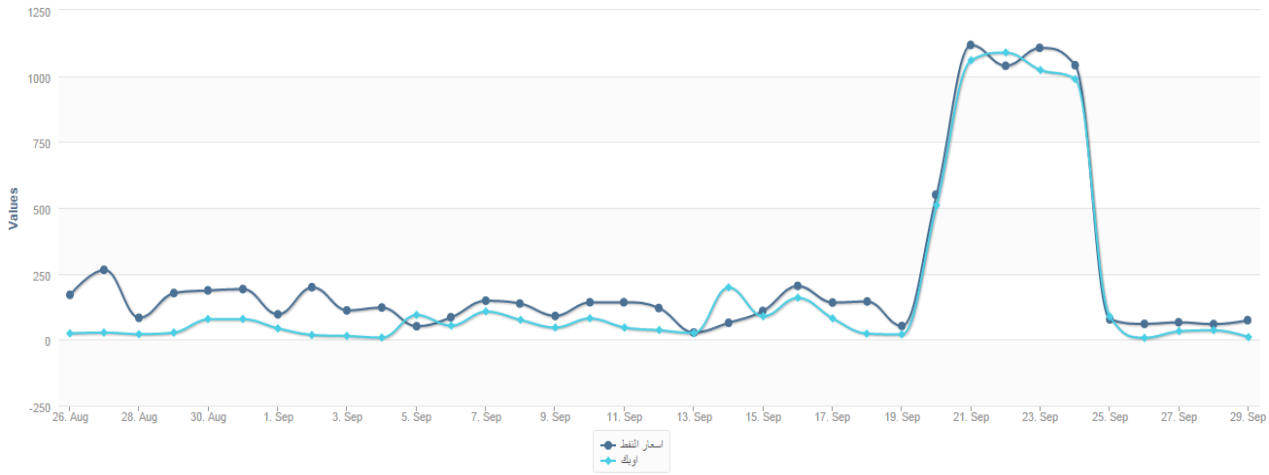

Figure 2. Timelines of "اسعار النفط , اوبك"

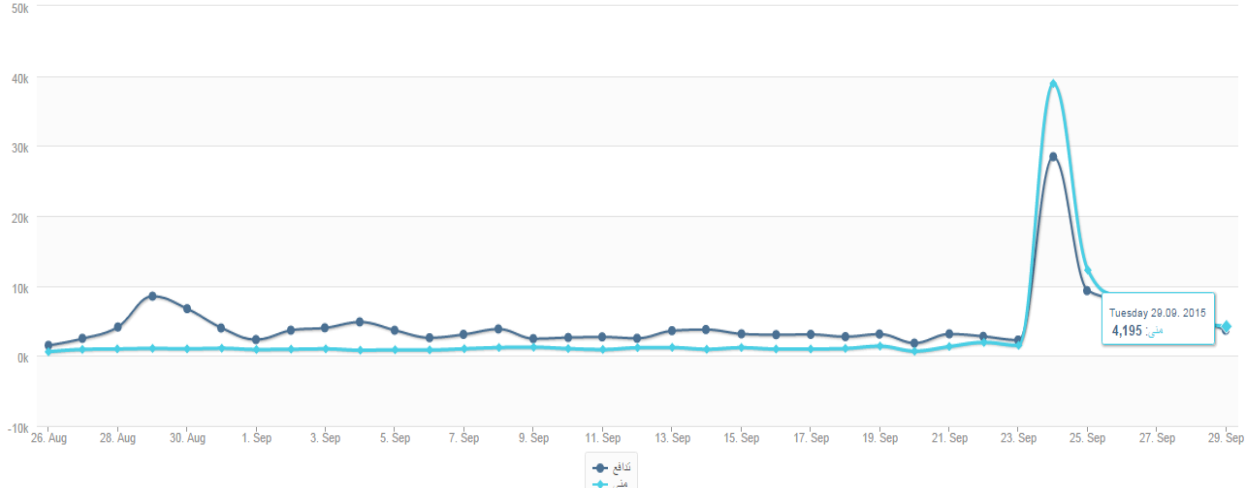

Figure 3. Timeline example (Term collocation)

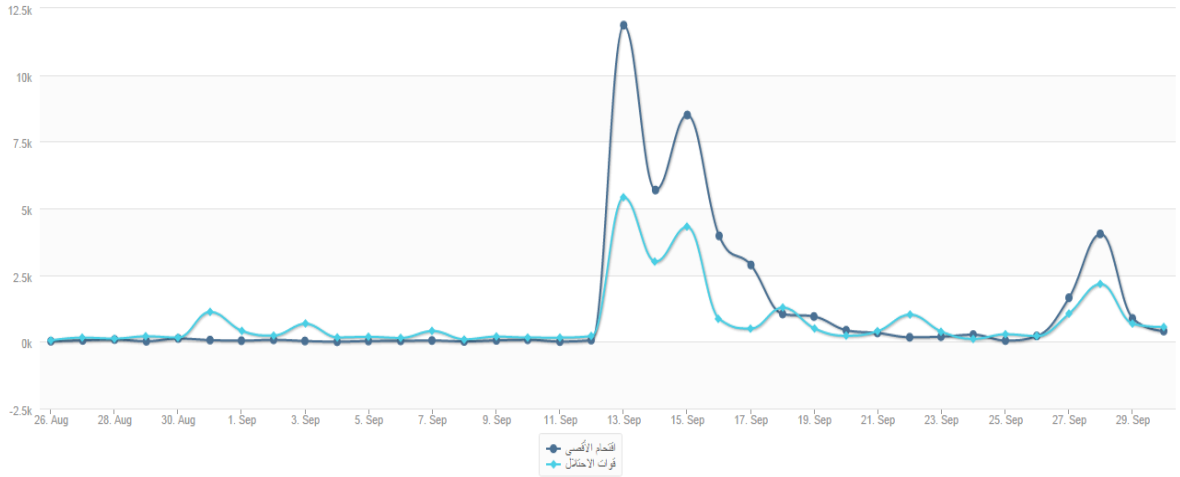

Figure 4. Event co-occurrence

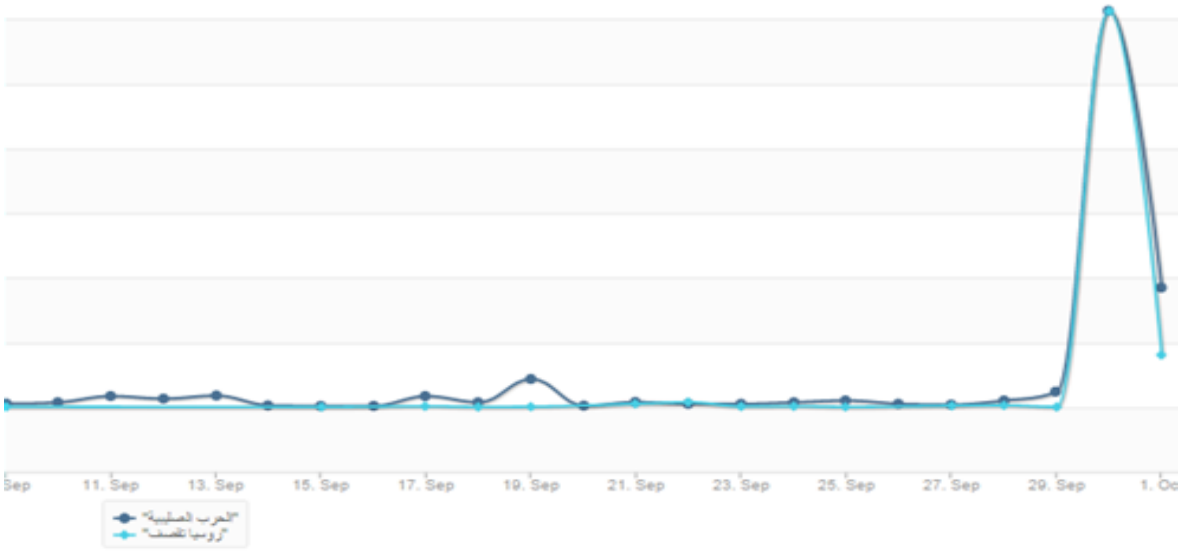

Figure 5. Community generated synonym

3.1 Time-series similarity search

Similarity search in timelines (time-series) is an interesting research direction to analyze stock prices data, weather forecast, biomedical measurements, etc. While there are several methods to find similarity between time series, the choice of a particular method is an application-dependent. Therefore, we are testing our hypothesis with a standard Dynamic Time Warping (Berndt and Clifford 1994) algorithm to measure the similarity between terms. There are several approaches that depend on the application. In our case the approach we need to use must consider the following assumptions:

1. Suppose that *t1* and *t2* are two timelines for two terms. *t1* and *t2* are similar if they have similar shapes. For example, figure (4, from 12/9 to 17/9) shows different frequencies between the two timelines. However, the shapes are similar.
2. Similar terms might not peak in the exact same day. *t1* could peak in a particular day and the other *t2* might peek in the next day. *t1* and *t2* are considered similar if they have similar peaks.
3. The presence of the peaks is more important that their magnitudes.

Dynamic time warping (DTW) is a technique that aligns two time series in which one time serie may be “warped” by stretching or shrinking its time axis. This alignment can be used to find corresponding regions or to determine the similarity between the two time series.

DTW focuses on aligning the peaks of the time lines without focusing on their magnitudes and it matches peaks even if they did not appear at the exact same time. This satisfies the assumptions mentioned above. DTW would consider *t1* and *t2* in figure (6) to be similar.

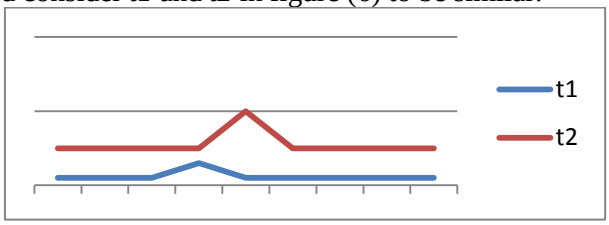

Figure 6. Two similar time series

3.2 DTW algorithm

DTW is a time series alignment algorithm that was originally used in voice recognition (Sakoe and Chiba 1978) It relates two time series of feature vectors by warping the time axis of one series onto another. Given two time series X and Y, Where:

$$X = x_1 + x_2 + x_3 + \ldots + x_i + \ldots + x_n$$
$$Y = y_1 + y_2 + y_3 + \ldots + y_i + \ldots + y_n$$

Algorithm 1 will produce the cost of aligning X and Y (warping them) the cost will be low if the two time series are similar.

```
int standardDWT(X, Y) {
// Where X = x1 + x2 + x3 + . . . + xi + . . . + xn and Y = y1 + y2 + y3 + . . . + yi + . . . + yn
   Create DTW[0..n, 0..m]
   Set the first row and column of DTW to infinity
   DTW[0, 0] = 0
   for i = 1 to n
      for j = 1 to m
         DTW[i, j] = d(X[i], Y[j])+ minimum(DTW[i-1, j] ,
                                            DTW[i  , j-1],
                                            DTW[i-1, j-1])
   return DTW[n, m]
}
```

Algorithm 1. Standard DWT

We start by filling a distance matrix DTW which has n × m elements; each element represents the warping distance between every two points in the time series. The warping distance between x_i and y_j is measured according to the following equation:

$$DTW(x_i, y_i) = d(x_i, y_i) + minimum(DTW(x_{i-1}, y_j) , DTW(x_i , y_{j-1}), DTW(x_{i-1}, y_{j-1}))$$

Where $d(x_i, y_i)$ is a distance function to calculate the distance between x_i and y_i. This version of DTW satisfies the monotonicity, continuity, boundary constrains demonstrated by (Sakoe and Chiba 1978, Keogh and Ratanamahatana 2004, Salvador and Chan 2007). We use the Euclidian distance as a distance function between x_i, y_i. So the distance will be calculated as follows:

$$d(x_i, y_i) = | x_i - y_i |$$

Frequency reading must be normalized to achieve meaningful results and to give more importance to peaks in relation to the average readings of a particular timeline. A frequency reading f is measured according to this equation:

$$Norm(f) = f - m$$

Where m is the average of frequencies for that term and the returned value from the algorithm indicates the cost of aligning the two normalized timelines. The similarity score described below indicates the possible similarity between the two timeline:

$$Similarity(X, Y) = 1 - cost/max(n, m)$$

Where *cost* is the returned value from the algorithm, n and m are the lengths of X and Y respectively. High similarity score means the probability that the two terms are related is high.

4 Data Collection

We are testing our approach with timelines collected by an online platform that addresses Arabic social media content and provides a platform to collect, search, monitor and analyze social media content. The platform has many functions. However, we are interested in the production of timelines which are archived through the following steps:

1. Data collection: Arabic tweets are collected using Twitter API. The online platform receives live feed from Twitter. Any non-Arabic tweets will be filtered.
2. Indexing: tweets are analyzed and indexed according to the terms they carry. Arabic analysis component is used for stemming and tokenization.

3. Reporting: the platform reports the frequencies for each term per time interval. Thus, we can build a timeline for each term.

The online system is available currently at "http://45.33.23.107". We are using its produced timelines and terms for our experiment.

5 Experimentation and Evaluation

Arabic tweets collected by the online platform during the month of May 2016 were analyzed. We selected 630 timelines for the most popular preterms in that month. Then we searched for similarities between them. The produced relationships were evaluated based on precision and recall.

The top 1108 relationships were rated by 2 evaluators (E1 and E2). Relationship between *t1* and *t2* is considered correct if the two evaluators found that *t1* and *t2* are event related or if they found that there is a terminological relationship (synonymy, acronym, hyponymy, antinomy, and hypernymy) between them. Using Cohen's Kappa coefficient (Cohen 1960) the inter-agreement score was 0.93 which indicates a substantial agreement between the evaluators.

5.1 Precision

We are trying to evaluate the precision of the similarity score according to this equation:

$$Precision = C_{th}/T_{th}$$

Where C_{th} is number of correct relationships with a score greater than the threshold *th*. T_{th} is total number of produced relationships with a score that is greater that *th*. When *th* is small the produced set of relationships increases but precision might decrease. When *th* = 0.85 the precision is 0.92. Figure (7) shows the precision in relation to the threshold.

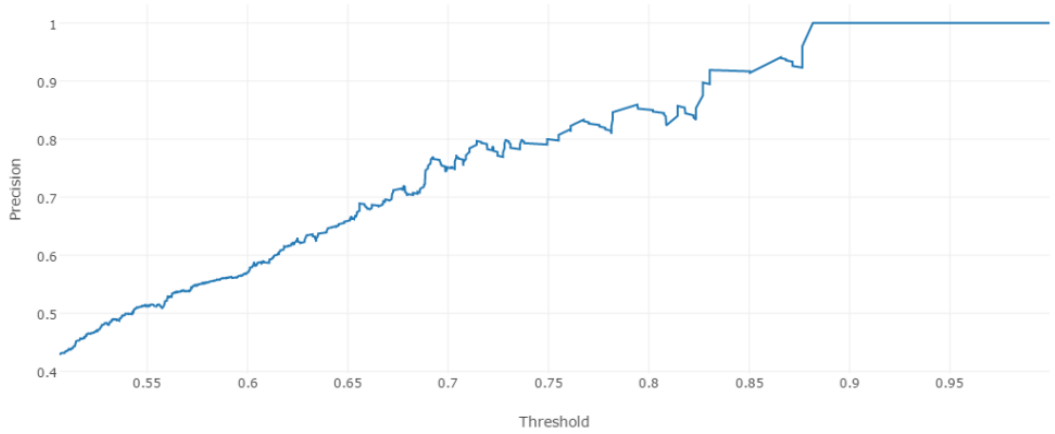

Figure 7. Precision

As you can see the precision starts to decline when *th* is below 0.5. The similarity score proved to be a good indicator of a relationship between preterms.

5.2 Recall

Recall is measured in terms of number of correct relationships extracted by our approach. When the threshold is 0.65 number of correct relationships is 200. Figure (8) shows the recall in relation to the threshold. When the threshold is 0.6 the precision is 0.61 and 270 correct relation were extracted from 630 preterm.

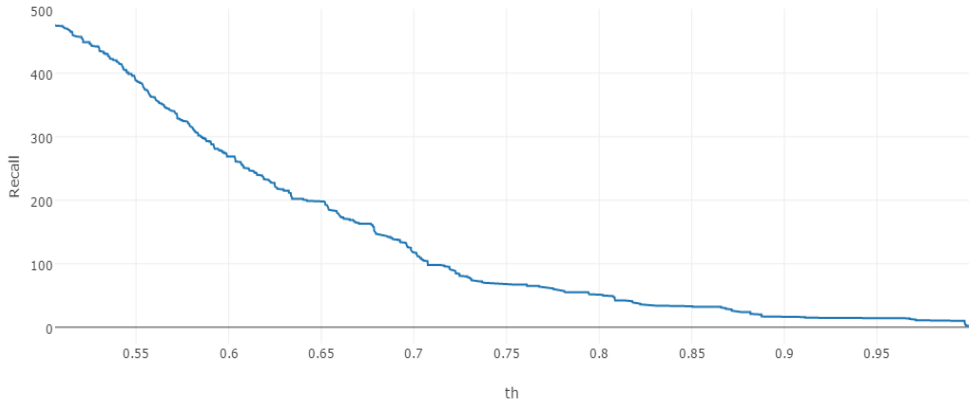

Figure 8. Recall

5.3 Assessment and sample results

Our approach has correctly identified terminological relationships between time sensitive preterms without analyzing the textual context; Table (2) shows sample results.

Table 2. Sample results

T1	T1 - English Translation	T2	T2 - English Translation	Similarity	Note
الثورة العربية الكبرى	The great Arab revolt	الثورة العربية	The Arab revolt	0.98	correct
ولي ولي العهد	Deputy crown prince	الرؤية السعودية 2030	Saudi vision 2030	0.96	correct
مئوية الثورة	Revolt Centennial	الثورة العربية الكبرى	The great Arab revolt	0.89	correct
الاحتلال الإسرائيلي	Israeli occupation	قوات الاحتلال	Occupation forces	0.88	correct
عيد الاستقلال	Independence day	الاعياد الوطنية	National holiday	0.88	correct
العراق	Iraq	الحشد الشعبي	Popular Mobilization Forces	0.83	correct
جرائم حرب	War Crimes	الحشد الشعبي	Popular Mobilization Forces	0.81	correct
الشرطة	Police	وزارة الداخلية	Ministry of interior affairs	0.8	correct
ولي العهد	Crown prince	خادم الحرمين	Custodian of the Two Holy Mosques	0.8	correct
ولي ولي العهد	Deputy crown prince	محمد بن سلمان	Mohammad bin Salman	0.8	correct
القصف الروسي	Russian bombing	العدوان الروسي	Russian aggression	0.76	correct
وزارة الصحة	Ministry of health	البترول	Petrol (oil)	0.73	incorrect
الثورة العربية الكبرى	The great Arab revolt	ولي ولي العهد	Deputy crown prince	0.7	incorrect

Extracting relationships between terms is a challenging task that needs large corpora, and specialists. The challenge increases when the terms are time sensitive Arabic terms. Our approach extracted 480 of relationships from 630 preterms with high precision; these relationships can be used in many applications, such as:

1. Extracted relationships can be post edited by specialists to enrich Arabic term bases.
2. Lexicon for social media analysis: auto microblogs classifications, auto tagging, sentiment analysis. In fact, we intend to use these relationships to dynamically extend a polarized lexicon for Arabic sentiment analysis.
3. These relationships can locate newly coined terms on an ontological resource.

The approach will be used on a larger scale to automatically discover related terms on-the-fly by analyzing online microblog feeds. The importance of this approach is that it does not rely on textual context; in fact many extracted relations were between terms that did not appear in the same tweet. Most of the wrongly extracted relationships were between key terms describing two separate events that took place at the same time. These errors can be reduced when the timeline is longer than 30 days.

6 Conclusions

We have presented an approach to extract terminological relationships between time-sensitive Arabic preterms. Our hypothesis is that terms that have similar history (timeline) are similar or related. We used Dynamic Time Warping algorithm to measure the similarity between terms. Our experiment produced 270 correct relationships out of 630 preterms with a precision of 0.61. The extracted information is crucial because it maps time-sensitive terms into a wider terminological map. The approach can be used to identify and connect terminology on-the-fly by analyzing microblogs feeds online, without relying on textual context (which is very limited in the case of online microblogs).

References

Agrawal, R., et al. (1993). Efficient Similarity Search In Sequence Databases. Proceedings of FODO. Illinois, USA.

Becker, H., et al. (2011). "Beyond trending topics: Real-world event identification on Twitter." In Proceedings of the Fifth International Conference on Weblogs and Social Media.

Berndt, D. J. and J. Clifford (1994). Using Dynamic Time Warping to Find Patterns in Time Series. KDD workshop, Seattle, WA.

Bothos, E., et al. (2010). "Using social media to predict future events with agent-based markets."

Cabre, M. T. and J. C. Sager (1999). "Terminology: Theory, methods, and applications." J. Benjamins Publishing: xii, 247 p.

Cabré, T. and R. Nazar (2011). Towards a new approach to the study of neology. Neology and Specialised Translation 4th Joint Seminar Organised by the CVC and Termisti.

Chen, H., et al. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly **36**(4): 1165-1188.

Cohen, J. (1960). "A coefficient of agreement for nominal scales." Educational and Psychological Measurement **20**(1): 37-46.

Cornolti, M., et al. (2013). A framework for benchmarking entity-annotation systems. Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee.

Daoud, D., et al. (2015). "Time-Sensitive Arabic Multiword Expressions Extraction from Social Networks." INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY.

Daoud, M. (2010). "Utilisation de ressources non conventionnelles et de méthodes contributives pour combler le fossé terminologique entre les langues en développant des" préterminologies" multilingues."

Daoud, M., et al. (2009). Constructing Multilingual Preterminological Graphs using various online-community resources. The Eighth International Symposium on Natural Language Processing (SNLP09), Thailand, Bangkok, IEEE.

Daoud, M., et al. (2009). "Constructing multilingual preterminological graphs using various online-community resources". the Eighth International Symposium on Natural Language Processing (SNLP2009), Thailand: pp. 116 - 121.

Daoud, M., et al. (2010). "Building Specialized Multilingual Lexical Graphs Using Community Resources." Lecture Notes in Computer Science - Resource Discovery **Volume 6162**: pp 94-109.

Daoud, M., et al. (2010). "Passive and Active Contribution to Multilingual Lexical Resources through Online Cultural Activities". NLPKE10, Beijing, China: 4 p.

Davidson, L. M. (1998). "Knowledge extraction technology for terminology."

Federmann, C., et al. (2012). Multilingual terminology acquisition for ontology-based information extraction. Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012).

Gallego Hernández, D. and S. Herrero Díaz (2014). "Terminology and French-Spanish business translation: evaluating terminology resources for the translation of accounting documents." MonTI : Monografías de Traducción e Interpretación **6**.

Gaussier, É. and J.-M. Langé (1997). "Some methods for the extraction of bilingual terminology." 1997). New Methods in Language Processing: 145-153.

Giannakidou, E., et al. (2014). Towards a Framework for Social Semiotic Mining. Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), ACM.

Grinev, M., et al. (2011). "Analytics for the RealTime Web." Proceedings of the VLDB Endowment **Volume 4**: pp 1391-1394.

Hartley, A. and C. Paris (1997). "Multilingual Document Production: from Support for Translating to Support for Authoring." Machine Translation 12 (1-2), pp. 109-29.

internetlivestats (2015). "Internet Live Stats - Internet Usage & Social Media Statistics." Retrieved 1/2/2015, 2015, from http://www.internetlivestats.com/one-second/.

Kageura, K. (2002). "The Dynamics of Terminology: A descriptive theory of term formation and terminological growth", Terminology and Lexicography Research and Practice 5, 322 p.

Kandias, M., et al. (2013). Proactive insider threat detection through social media: The YouTube case. Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society, ACM.

Keogh, E. and C. Ratanamahatana (2004). "Exact indexing of dynamic time warping." Knowledge and Information Systems: 358-386.

Kim, Y. G., et al. (2005). "Terminology construction workflow for Korean-English patent MT". MT Summit X. Phuket, Thailand: 5 p.

Kwak, H., et al. (2010). What is Twitter, a Social Network or a News Media? The 19th International World Wide Web (WWW) Conference, Raleigh NC (USA).

Peñas, A., et al. (2001). Corpus-based terminology extraction applied to information access. Proceedings of Corpus Linguistics, Citeseer.

Roche, C., et al. (2009). Ontoterminology: A new paradigm for terminology. International Conference on Knowledge Engineering and Ontology Development.

Sager, J. C. (1990). A Practical Course in Terminology Processing Amesterdam/Philadelphia, John Benjamins Publishing Company, 266 p.

Sakoe, H. and S. Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition." IEEE Transactions on Acoustics, Speech and Signal Processing: 43- 49.

Salvador, S. and P. Chan (2007). "Toward accurate dynamic time warping in linear time and space." Intelligent Data Analysis: 561 - 580.

Speriosu, M., et al. (2011). "Twitter polarity classification with label propagation over lexical links and the follower graph." EMNLP '11 Proceedings of the First Workshop on Unsupervised Learning in NLP: 53-63

Suonuuti, H. (1997). " Guide to Terminology." Nordterm 8. Helsinki:TSK.

Twitter (2015). "Twitter." Retrieved 1/2/2015, 2015, from twitter.com.

Uherčík, T., et al. (2013). "Utilizing Microblogs for Web Page Relevant Term Acquisition." SOFSEM 2013: Theory and Practice of Computer Science - Lecture Notes in Computer Science **Volume 7741**: pp 457-468.

Vasconcellos, M., et al. (2001). "Terminology and machine translation." Handbook of terminology management **2**: 697-723.

Yang, Y. T., et al. (2013). "Mining social media and web searches for disease detection." Journal of public health research **2**(1): 17.

Zhao, W. X., et al. (2011). "Topical keyphrase extraction from Twitter." HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 **379-388**.