Linguistic Resources & Topic Models for the Analysis of Persian Poems

Ehsaneddin Asgari and Jean-Cédric Chappelier
Ecole Polytechnique Fédérale de Lausanne (EPFL)
School of Computer and Communication Sciences (IC)
CH-1015 Lausanne ; Switzerland
ehsaneddin.asgari@epfl.ch and jean-cedric.chappelier@epfl.ch

Abstract

This paper describes the usage of Natural Language Processing tools, mostly probabilistic topic modeling, to study semantics (word correlations) in a collection of Persian poems consisting of roughly 18k poems from 30 different poets. For this study, we put a lot of effort in the preprocessing and the development of a large scope lexicon supporting both modern and ancient Persian. In the analysis step, we obtained very interesting and meaningful results regarding the correlation between poets and topics, their evolution through time, as well as the correlation between the topics and the metre used in the poems. This work should thus provide valuable results to literature researchers, especially for those working on stylistics or comparative literature.

1 Context and Objectives

The purpose of this work is to use Natural Language Processing (NLP) tools, among which probabilistic topic models (Buntine, 2002; Blei et al., 2003; Blei, 2012), to study word correlations in a special type of Persian poems called “Ghazal” (غزل), one of the most popular Persian poem forms originating in 6th Arabic century.

Ghazal is a poetic form consisting of rhythmic couplets with a rhyming refrain (see Figure 1). Each couplet consists of two phrases, called hemistichs. Syllables in all of the hemistichs of a given Ghazal follow the same pattern of heavy and light syllables. Such a pattern introduces a musical rhythm, called *metre*. Metre is one of the most important properties of Persian poems and the reason why usual Persian grammar rules can be violated in poems, especially the order of the parts of speech. There exist about 300 metres in Persian poems, 270 of which are rare, the vast majority of poems composed only from 30 metres (Mojiry and Minaei-Bidgoli, 2008).

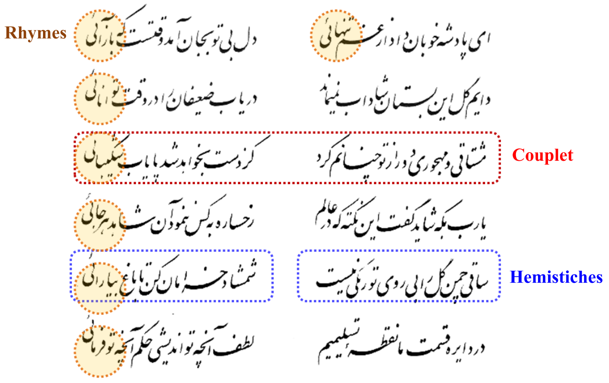

Figure 1: Elements of a typical Ghazal (by Hafez, calligraphed by K. Khoroush). Note that Persian is right to left in writing.

Ghazal traditionally deals with just one subject, each couplet focusing on one idea. The words in a couplet are thus very correlated. However, depending on the rest of the couplets, the message of a couplet could often be interpreted differently due to the many literature techniques that can be found in Ghazals, e.g. metaphors, homonyms, personification, paradox, alliteration.

For this study, we downloaded from the Ganjoor poems website[1], with free permission to use, a Ghazal collection corresponding to 30 poets, from Hakim Sanai (1080) to Rahi Moayyeri (1968), with a total of 17,939 Ghazals containing about 170,000 couplets. The metres, as determined by experts (Shamisa, 2004), are also provided for most poems.

[1] http://ganjoor.net/.

Proceedings of the Second Workshop on Computational Linguistics for Literature, pages 23–31,
Atlanta, Georgia, June 14, 2013. ©2013 Association for Computational Linguistics

We put a lot of effort into the preprocessing, so as to provide more informative input to the modeling step. For this, we built a lexicon supporting both modern and ancient Persian, as explained in Section 2. In addition, we developed several preprocessing tools for Persian and adapted them to poems, as detailed in Section 3. In the analysis step, exploiting Probabilistic Topic Models (Blei, 2012), promising results were obtained as described in Section 4: strong correlation between poets and topics was found by the model, as well as relevant patterns in the dynamics of the topics over years; good correlation between topics and poem metre was also observed.

2 Modern and Ancient Persian Lexicon

This section presents the Persian lexicon we built, which supports both modern and ancient Persian words and morphology and provides lemmas for all forms. This lexicon could thus be useful to many research projects related to both traditional and modern Persian text processing. Its total size is about 1.8 million terms, including the online version[2] of the largest Persian Dictionary today (Dehkhoda, 1963). This is quite large in comparison with e.g. the morphological lexicon provided by Sagot & Walther (2010), of about $600k$ terms in total.

2.1 Verbs

Taking advantage of the verb root collection provided by Dadegan group (Rasooli et al., 2011), we conjugated all of the regular forms of the Persian verbs which exist in modern Persian using grammars provided by M. R. Bateni (1970), and added them with their root forms (lemmas) to the lexicon. We also added ancient grammatical forms, referring to ancient grammar books for Persian (Bateni, 1970; P. N. Xanlāri, 2009).

Persian verb conjugation seems to be simple: normally each verb has two roots, past and present. In each conjugated form, the corresponding root comes with some prefixes and attached pronouns in a predefined order. However, phonological rules introduce some difficulties through so-called *mediators*. For instance, the verb آراستن (**ârâstan**, meaning ”*to decorate*” or ”*to attire*”) has آرا (**ârâ**) as present root and آراست (**ârâst**) as past root. Its injunctive form requires it to be preceded by بِـ (**be**), leading to بارا (**beârâ**). However, according to phonological rules, when a consonant attaches to آ (**â**), a یـ (**y**) is introduced as a mediator. So the correct injunctive form is بیارا (**byârâ**, “*decorate!*”).

Mediators occur mainly when a consonant comes before **â** or when a syllable comes after **â** or و (**u**). But the problem is slightly more complicated. For instance, the present verb for جستن (**jostan**, “*seeking*”) is جو (**ju**). Thus when the pronoun ـَم (**am**, “*I*”) is attached, the conjugated form should be جویم (**juyam**, “*I seek*”), with a mediator. However, the root **ju** has also a homograph **jav** (also written جو) which is the present root of جویدن (**javidan**, “*chewing*”). Since here و is pronounced **v**, not **u**, there is no need for a mediator and the final form is جوم (**javam**, “*I chew*”). Therefore, naively applying the above mentioned simple rules is wrong and we must proceed more specifically. To overcome this kind of problem, we studied the related verbs one by one and introduced the necessary exceptions.

In poems, things are becoming even more complicated. Since metre and rhyme are really key parts of the poem, poets sometimes waives the regular structures and rules in order to save the rhyme or the metre (Tabib, 2005). For instance, F. Araqi in one of his Ghazals decided to use the verb form می‌نایی (**mi-nâyi**, “*you are not coming*”) which does not follow the mediator rules, as it must be می‌نیایی (**mi-naâyayi**). The poet decided to use the above form, which still makes sense, to preserve the metre.

The problem of mediators aside, the orders of parts in the verb structures are also sometimes changed to preserve the metre/rhyme. For instance in the future tense, the compound part of compound verbs has normally to come first. A concrete example is given by the verb جان خواهد سپرد (**jân xâhad sepord** means “*(s)he will give up his spirit and will die*”), which is written by Hafez as: خواهد سپرد جان (**xâhad sepord jân**). To tackle these variations, we included in our lexicon all the alternative forms mentioned by Tabib (2005).

As already mentioned, the considered poem collection ranges from 1080 to 1968. From a linguistics point of view some grammatical structures of the language have changed over this long period of time. For instance, in ancient Persian the prefix for

[2] `http://www.loghatnaameh.org/`.

the continuity of verb was همی (**hami**); today only می (**mi**) is used. Many kinds of changes could be observed when ancient grammars are compared to the modern one. The relevant structures to the mentioned period of time were extracted from a grammar book of ancient Persian (P. N. Xanlāri, 2009) and included in our lexicon.

Starting from the 4,162 infinitives provided by Dadegan group (Rasooli et al., 2011) and considering ancient grammars, mediators, and properties of poetic forms, we ended up with about 1.6 million different conjugated verb forms. The underlying new structures have exhaustively been tested by a native Persian graduate student in literature and linguistics. This validation took about one hundred hours of work, spot-checking all the conjugations for random selected infinitives.

2.2 Other words (than verbs)

The verbs aside, we also needed a complete list of other words. The existing usual Persian electronic lexica were insufficient for our purpose because they are mainly based on newspapers and do not necessarily support ancient words. For our purpose, the ongoing effort of Dehkhoda Online Dictionary[3] looked promising. Dehkhoda dictionary (Dehkhoda, 1963) is the largest comprehensive Persian dictionary ever published, comprising 16 volumes (more than 27,000 pages), entailing over 45 years of efforts by Aliakbar Dehkhoda and other experts and it is still ongoing. The Dehkhoda Online Dictionary Council fortunately approved our request to use their work which currently contains 343,466 entries (for 234,425 distinct forms).

Besides the Dehkhoda Online Dictionary, we added the free Virastyar Persian lexicon[4]. Although the is size is one tenth of Dehkhoda's, it contains several new imported words, not found in Dehkhoda. All together, we ended up with a lexicon of 246,850 distinct surface forms. For each surface form, we also provide the list of corresponding roots (lemmas).

3 Preprocessing

Preprocessing is an essential part in NLP which usually plays an important role in the overall performance of the system. In this work, preprocessing for Persian Ghazals consists of tokenization, normalization, stemming/lemmatization and filtering.

3.1 Tokenization

The purpose of tokenization is to split the poems into word/token sequences. As an illustration, a hemistich like

شاه شمشاد قدان خسرو شیرین دهنان

is split into the following tokens:

شاه / شمشاد / قدان / خسرو / شیرین / دهنان.

The tokenization was done using separator characters like white spaces, punctuation, etc. However, half-spaces made this process quite complicated, as most of them appeared to be ambiguous.

Half-space is a hidden character which avoids preceding letters to be attached to the following letters; the letters in Persian having different glyphs when attached to the preceding letters or not.

For instance, می‌رفت (**mi-raft**, "*was going*"), here written with a half-space separating its two parts, **mi** (می) and **raft** (رفت) would be written میرفت without the half-space (notice the difference in the middle).

Half-spaces carry useful information, e.g. for recognizing compound words. However, they were not reliable enough in the poem collection used.

The main challenges we had to face related to half-spaces were related to continuous verbs. In Persian, continuous verbs have a prefix **mi** (می)which should be separated from the rest of the verb by a half-space. However, it was sometimes written using full-spaces and sometimes even without any space at all. For instance **mi-goft** ("*was saying*") should be written with a half-space: می‌گفت but was sometimes written using a full space: می گفت, and even sometimes without any separator: میگفت. The problem of identifying continuous verbs is even more complicated in poems because the prefix (**mi**) is the homograph of a word meaning "*wine*" (**mey:** می), quite frequent in Persian poems.

For dealing with continuous verbs, we apply the following heuristic: in the structure of continuous verbs, the prefix **mi** comes before the root of verbs, thus, if a root of a verb comes just after a **mi**, then we

[3] http://www.loghatnaameh.org/.

[4] http://www.virastyar.ir/data/.

can consider it as a continuous verb. However, many **mi**'s meaning *wine* would be considered as prefixes using this too simple heuristic, because the most frequent letter in Persian آ (**â**) is also a verb root. For instance, in phrase **mey-e-âsemâni:** می آسمانی , **mey** means "*wine*" and the second part آسمانی means "*related to heaven*" (as an adjective, not a verb). To consider **mi** as a prefix, we thus constrained the token after it to start with a root longer than 2 letters.

The mentioned rule improves the process of tokenization. However, there are still some cases which are really complicated even for a human to decide. For instance, **mi-âlud:** می آلود ("*was polluting*") and **mey-âlud:** می آلود ("*polluted with wine*") are homographs in Persian; whose precise tokenization requires contextual information or even metre to decide which one is more suited. As a simple solution we can consider **mey-âlud** and any other known compound forms of **mey** as compound words and add them to our lexicon. Taking the advantages of this solution for such ambiguous words, we can identify if there is any ambiguity and given that there is some, we can pass all of them to the next processing steps, not deciding too soon.

Continuous verbs aside, present perfect verbs, prefix verbs, and compound verbs have also two parts which might be separated with half-space or full white-space. For instance, **rafteh-am** ("*have gone*") might appear with a half-space: رفته‌ام , without any separator: رفتهام , or with a full space: رفته ام.

Since the tokenization was complicated and requires linguistic knowledge, especially to properly handle half-spaces, we designed it in two steps: first a basic version to bootstrap the process before character normalization (next subsection), and later a refinement of the basic tokenization, taking advantage of the rich Persian lexicon we built.

As a first tokenization step, we took the advantage of the fact that the number of tokens in a hemistich is intuitively between four and ten, because of Ghazals' metre rules. We thus decided that when full-space tokenization had less than four tokens, then both full- and half-spaces must be considered as separators. If the number of tokens obtained this way is more than four, the tokenization is forwarded to the next step. Otherwise, if there is still less than four tokens, the hemistich is marked for manual checking. The number of hemistichs that required manual fixation was very low, about 40 out of 340,000.

3.2 Normalization

In Persian fonts, several letters have more than one form, because of different writing style related to different pronunciations; for instance **âmrika**: آمریکا, **emrika**: امریکا ("*America*"); and of different characters encoding of Arabic letters; for instance **anâr** ("*pomegranate*") might be written انار or أنار.

We get rid of these meaningless variations by normalizing the characters. This normalization has to come *after* basic tokenization because of the unreliable half-spaces, to be handled first, that interfere with the written form of some letters.

We first used *both* Arabic and Persian normalizers of Lucene[5]: in the Persian version, most of the cases are considered except different alefs (first letter of Persian alphabet), which are properly handled by the Arabic normalizer. We furthermore added the following rules to Lucene modules:

- Normalization for **vâv** and **ye**:

  There are two different forms of **vâv**: وor ؤ , which is rather Arabic, not preferred in Persian. For instance, word **mo'men** ("*believer*") could be written مومن or مؤمن.

  We have a similar case with **ye** which might be written ئـ or یـ. For instance, **âyine** ("*mirror*") might be written آیینه or آئینه.

- Some characters exist which are optional in Persian writing for instance light vowels, **tašdid** (sign of emphasis: ـّ in محمّد), and **tanvin**s, three signs could be attached at the end of some words, e.g. ـً in حضوراً . Some of them were implemented in Lucene Arabic normalizer, some in the Persian normalizer and some in none of them.

- Removal of the optional **hamze** sign ء at the end of word, for instance: املاء.

- Removal (without any change in the meaning) of some Arabic characters that do not normally appear in Persian but were present in the corpus, e.g. ٍ (**tanvin kasre**), ٌ (**tanvin zamme**).

[5] `http://lucene.apache.org/`.

- Removal (without any change in the meaning) of adornment (calligraphy) characters, e.g. dashes, ْ(**sokun**), and ٓ(**mad**).

As explained in the former subsection, the final tokenization was postponed due to the difficult ambiguities introduced by half-/full-space confusions. To finalized it after character normalization, taking the advantage of our lexicon, we considered all bigrams, trigrams and 4-grams of tokens obtained and checked whether they correspond to a valid form in the lexicon. Out of 2,876,929 tokens, we had 330,644 (valid) bigrams, 12,973 trigrams and 386 4-grams.

3.3 Stemming/Lemmatization

The purpose of stemming/lemmatization[6] is to regroup (using the same writing) words of similar root, in order to reduce (hopefully the non-significant part of) the variability of the documents processed.

Although a free Persian stemmer PerStem exists (Jadidinejad et al., 2009)[7], its limitations we observed (see below) encouraged us to build our own stemmer.

Since Persian is an affixive language, lemmatization is achieved by removing plural signs, attached pronouns, prefixes and suffixes to obtain the root. We thus collected a list of these and enriched it using affixes provided by Adib Tousi (1974) and by Tabtabai (2007). Then we designed a flowchart to iteratively remove the unwanted parts from the normalized token until we get a simple word contained in the lexicon or a word with a length of less than 4 letters. The experiences showed us it is more appropriate to remove prefixes first, then suffixes. Even in suffix removal, the removal order is a crucial issue. Since some words have more than one suffix and the set of suffixes is not a prefix-free set, a wrong removal order can leads to removing a wrong suffix and might result in finishing the removal too early, where there still exist some letters to be removed. For instance, the word کتابهایشان (**ketâbhâyešân**, "*their books*") should be reduced to کتاب (**ketâb**, "*book*"). It has three suffixes ها (**hâ**, plural marker), ی (**ye**, mediator) and شان (**šan**, "*their*" as a attached pronoun). However, **šan** has two prefixes which are also suffixes: ن (**N**, infinitive mark) and ان (**ân**, plural mark for nouns). Such cases are not considered in PerStem, and the affixes removal is stopped too early. In order to overcome this problem in our stemmer, we generated all of the possible combinations of affixes and add them to our affixes collection. Then the largest possible one is removed from the token at each step.

We then checked for the term in the lexicon and return its lemmas when matched. If we could not find any matched term in the lexicon, we manually check the token. Doing so, we realized that because of the missing/wrong spaces, most of these tokens wrongly attached to conjunctions. For this specific purpose, we partially modified the list of affixes and applied the stemmer again on these out of vocabulary forms, ending up with the proper information.

In the case of homographs, for instance نشستی that could be read as **nešasti** ("*you sat*") or as **našosti** ("*you did not wash*"), we pass *all* possible interpretations to the next processing step. For instance, the result of the lemmatization of نشستی is "'to sit' or 'to wash'", i.e. both lemmas.

3.4 Filtering

In order to reduce even further the input variability, some filtering has been performed based both on frequencies and on a standard list of "stop-words", some extremely common words which are normally meaningless (at least independently).

The general strategy for determining stop-words is to sort the terms by their frequencies in the collection, consider the most frequent ones and then filter them manually with respect to the domain. Doing so, we found stop-words well suited for the poem collection considered, which is slightly different from stop-words in normal Persian text (poem specific, and typographical error occurred in the corpus used). We also combined this set with a (manually chosen) subset of stop-words provided by K. Taghva (2003).

4 Topic Modeling

After preprocessing, we studied the correlations among words in Ghazals using "probabilistic topic

[6] Stemming reduces words to their stems, using rather crude algorithms and basic morphological rules, while lemmatization uses more advanced morphological analysis and lexical resources to find the root form, named lemma.

[7] `http://www.ling.ohio-state.edu/~jonsafari/persian_nlp.html`.

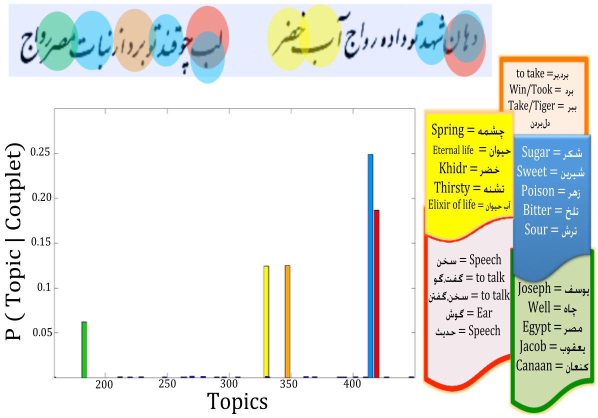

Figure 2: Probabilistic distribution over the topics (learned in an unsupervised manner) for one specific couplet: the horizontal axis stands for the topics and the vertical axis for the probability of each topic for the couplet considered. Notice how only a few topics are used in the couplet. The most probable words for the five most probable topics for this couplet are also provided on the right. On top, an example of a possible assignment of these topics to the words in the couplet considered is provided. Each color represents one of the 5 most probable topics.

models" (Buntine, 2002; Blei, 2012), more precisely Latent Dirichlet Allocation (LDA) (Blei et al., 2003)[8]. We looked into correlation between topics and poets, as well as between topics and metres, and obtained very interesting results.

4.1 Model

Probabilistic topic models are unsupervised generative models which represent documents as mixtures of *topics*, rather than (only) collections of terms (Blei, 2012). "Topics" are nothing else but probability distributions over the vocabulary that are learned in an unsupervised manner. Probabilistic topic models allow us to represent documents at a higher level (topics rather than words) with much fewer parameters. A typical example is given in Figure 2.

Taking advantage from conditional co-occurrences through topics, these models are able to take both polysemy and synonymy into account. To illustrate how such models behave, we could for instance consider the polysemic term

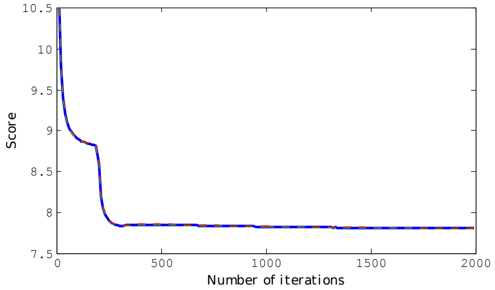

Figure 3: Learning score w.r.t number of iterations. After the iteration 200, hyper-parameter optimization starts and around 600 the score has converged. ±1-standard-deviation curves determined using 1x10-fold cross validation cannot be distinguished from the average curve.

شیرین (**širin/Shirin**, meaning "*sweet*" but also standing for the name of a well-known woman from a famous story), which appeared in the ten most frequent terms of topics 413 and 337 (blue words in Table 1). Two topics presented in Table 1 are showing different contexts that can include **širin** as a keyword. Topic 413 appeared to refer to contexts related to sweetness, whereas topic 337 appeared to refer to a famous Persian tragic romance, "*Khosrow and Shirin*", a.k.a. "*Shirin and Farhad*".

Furthermore, since ambiguity (homonymy) is a literature technique, sometimes poets use **širin** somewhere that can refer to both contexts. That could be the reason why شکر (**šekar**, "*sugar*"), represented in green, appears in frequent terms of both topics.

One key issue using these kind of models regards the choice of the number of topics. To decide the appropriate number, we measured the model quality with held-out log-likelihood (estimated on validation set) using 1x10-fold cross validation (Wallach et al., 2009; Buntine, 2009).[9] We ran each fold for 2000 iterations (convergence checked; see Figure 3) doing hyper-parameter optimization (Wallach et al., 2010) after 200 iterations. We observe that the log-likelihood decreases, and stabilizes around 400/500 topics (see Figure 4). We thus considered 500 topics to be a good order of magnitude for this corpus.

[8] We used Mallet software, version 2.0.7; `http://mallet.cs.umass.edu/`.

[9] Note that the evaluation method implemented in Mallet is the biased method provided by Wallach (2009) and not the proper methods suggested by Buntine (2009).

Table 1: 10 Most probable terms chosen from three topics (among 500 topics).

Topic 290	Topic 413	Topic 337
candle =شمع	sugar =شکر	Shirin =شیرین
butterfly =پروانه	sweet =شیرین	Farhad =فرهاد
light =چراغ	poison =زهر	Khosrow =خسرو
to tear =درید،در	bitter =تلخ	mountain =کوه
to burn =سوخت،سوز	sour =ترش	to carve or to do =کندن or کردن
bright =روشن	sugar =قند	sweet life =شیرین جان
society =انجمن	mouth =دهان	mount cutting =کن کوه
clique =محفل	honey =شهد	axe =تیشه
fire =اتش	palate =کام	blessed =خوبان
flame =شعله	bitterness =تلخی	sugar =شکر

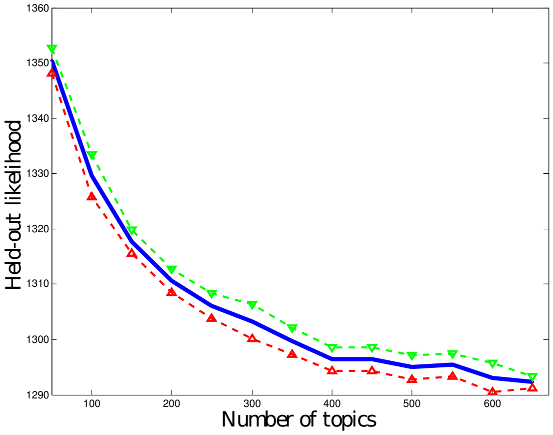

Figure 4: Held-out log-likelihood versus number of topics. ±1-stand.-dev. curves obtained by 1x10-fold cross-validation are also shown (dashed lines).

4.2 Correlation between Topics and Poets

Good correlation between some topics and poets has been observed. To investigate this correlation further, the joined probability of topics and poets is measured and the results are shown in Figure 5. It can be observed that there is a strong correlation between poets and topics. Some general topics (used by all the poets) also appear (as vertical darker lines).

Another good illustration of this correlation is given in Figure 6 which illustrates the proportions of four different topics for the 30 poets ordered by their lifetime. Some relevant patterns can be observed. For instance, the topic related to "Joseph" (blue) and the one related to "Mirror" (violet) are correlated. In Persian literature, Joseph is the symbol of beauty and

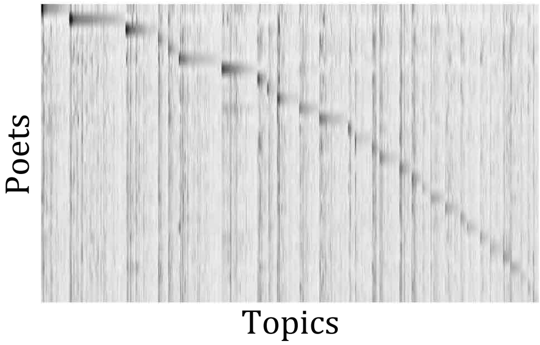

Figure 5: Correlation between (automatically found) topics and poets: the joint probability $P(\text{topic, poet})$ is plotted in dark shades; the darker the shade, the higher the probability. The dark mark along the diagonal thus illustrates a very good correlation (conditional probability, in fact) between topics and poets. For a better visualization, both rows and columns have here been reordered.

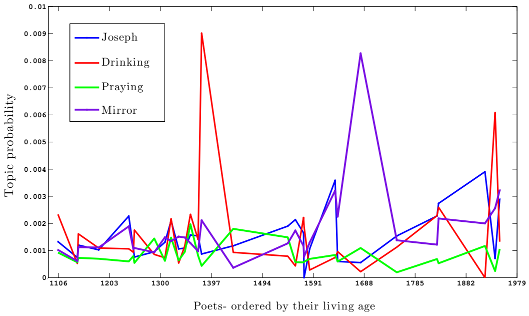

Figure 6: The probability of four different (automatically found) topics over the time. X-axis shows the middle of lifetime of the poets.

beauty can be perceived by means of the mirror. This is the reason why these two topics are somehow correlated. Moreover, the “Mirror” topic has an independent peak around 1700 A.D. This corresponds to Bidel Dehlave, so-called ”poet of mirrors” (Kadkani, 2007), who very often refers to mirrors in his poems.

Another pattern relates to drinking, which in Persian mystical literature refers to a grace from heaven. The main era of mystical literature is between 1300 and 1400 AD. As it can be observed from Figure 6, “Praying” and “Drinking” topics have similar curves in this time period, as expected. The independent peak corresponds to the poet Awhadi Maraghai who uses words related to drinking very much.

4.3 Correlation between Topics and Metre

There is supposed to be a relation between the happiness or sadness of the words in a poem and its melody (metre). Vahidian Kamyar (Kamyar, 1985), for instance, provides a list of metres and their corresponded feeling.

We thus also wanted to investigated whether there was any correlation between the metres and the topics learned in an unsupervised manner. To answer this question, we encoded the 30 metres provided in the original corpus as a (new random) term each, and then added the corresponding “metre term” once to each couplet. Then a topic model has been estimated.

The results obtained confirmed Kamyar’s observations. For instance, the topics that have as probable term the “metre term” corresponding to the metre Kamyar associates to requiem, parting, pain, regret and complain (فعلن فعلاتن فعلاتن) are presented in Table 2. As you can see all of the three topics presented are showing a kind of sadness.

Table 2: 8 Most probable terms chosen from three topics related to a metre usually related to sadness.

Topic 43 ($\simeq$ “Suffering”)	
رنجید،رنج =to suffer	راحت =comfort
رنج =pain	شفا =healing
بيمار =patient	رنجور =ill
طبيب =doctor	بيماري =illness

Topic 154 ($\simeq$ “Crying”)	
اشك =tear	سيل =flood
چكيد،چك =to trickle	روان =fluid
مژه =eyelash	اشكم =my tear
گريه =cry	سرشك =drop

Topic 279 ($\simeq$ “Love and Burn”)	
سوخت،سوز =to burn	شمع =candle
سوخته =burned (adj.)	عشق =love
اتش =fire	عود =oud ($\simeq$ guitar)
سوخت =burned or fuel (N.)	جگر =liver ($\simeq$ heart)

“Love & Burn” topic is not surprising for people used to Persian poetry as the butterfly—candle metaphor is often used، reminding of a common belief among Persians that butterflies love candles to the ultimate level of love so as to vanish in the presence of candle by burning in its fire.

5 Conclusion

With this study, we show that we can fruitfully analyze Persian poems, both for modern and ancient Persian, using NLP tools. This was not a priori obvious due to their specific nature (special form, ancient vocabulary and grammar, ...).

We put a lot of effort into the preprocessing, adapting it to poems, and in the development of a large scope lexicon supporting both modern and ancient Persian. In the analysis step, exploiting the power of probabilistic topic models, we obtained very interesting and meaningful results. We found strong correlation between poets and topics, as well as relevant patterns in the dynamics of topics over years. Correlation between the topics present in the poems and their metre was also observed.

As far as we know, this study is the first semantic study of Persian poems from a computational point of view. It provides valuable results for literature researchers, specially for those working in stylistics.

Follow-up work will include building a semantic search tool and a poem recommender system.

Acknowledgments The authors would like to warmly thank Mrs. Fereshteh Jafari for her help with the validation of the verb lexicon as well as the anonymous reviewers for their helpful comments.

References

[Bateni1970] M. R. Bateni. 1970. *The description of grammar structure in Persian language (توصیف ساختمان دستوری زبان فارسی)*. AmirKabir, Tehran.

[Blei et al.2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January.

[Blei2012] D. M. Blei. 2012. Probabilitic topic models. *Communications of the ACM*, 55(4):77–84, April.

[Buntine2002] W. Buntine. 2002. Variational extensions to EM and multinomial PCA. In *Proc. of ECML'02*, volume 2430 of *LNAI*, pages 23–34.

[Buntine2009] W. Buntine. 2009. Estimating likelihoods for topic models. In *Proc. of ACML'09*, volume 5828 of *LNAI*, pages 51–64.

[Dehkhoda1963] A.-A. Dehkhoda, editor. 1963. *The Dehkhoda Dictionary*. Tehran University Press.

[Jadidinejad et al.2009] A. H. Jadidinejad, F. Mahmoudi, and J. Dehdari. 2009. Evaluation of PerStem: a simple and efficient stemming algorithm for persian. In *Proc. 10th Cross-Language Evaluation Forum Conf. (CLEF'09)*, pages 98–101. Springer-Verlag.

[Kadkani2007] M. R. Shafiee Kadkani. 2007. *Poet of mirrors (شاعر آیینه‌ها)*. Agaah.

[Kamyar1985] T. Vahidan Kamyar. 1985. Metre in persian poems (اوزان ایقاعی شعر فارسی). Technical report, Department of Literature and Human Sciences, Ferdowsi University of Mashhad.

[Mojiry and Minaei-Bidgoli2008] M. M. Mojiry and B. Minaei-Bidgoli. 2008. Persian poem rhythm recognition: A new application of text mining. In *Proc. of IDMC'08*, Amir Kabir University.

[P. N. Xanl□ari2009] E. Mostasharniya P. N. Xanl□ari. 2009. *The Historical Grammar of Persian Language (دستور تاریخی زبان فارسی)*. Tose'eye Iran, 7th edition. (1st edititon: 1995).

[Rasooli et al.2011] M. S. Rasooli, A. Moloodi, M. Kouhestani, and B. MinaeiBidgoli. 2011. A syntactic valency lexicon for persian verbs: The first steps towards persian dependency treebank. In 5th *Language & Technology Conference (LTC): Human Language Technologies: a Challenger for Computer Science and Linguistics*.

[Sagot and Walther2010] B. Sagot and G. Walther. 2010. A morphological lexicon for the persian language. In *Proc. of the* 7th *Conf. on Int. Language Resources and Evaluation (LREC'10)*, pages 300–303.

[Shamisa2004] S. Shamisa. 2004. *An introduction to prosody (آشنایی با قافیه و عروض)*. Mitra, 4th edition.

[Tabatabai2007] A. Tabatabai. 2007. Persian language etymology (صرف زبان فارسی). *Bokhara Magazine*, 63:212–242, November.

[Tabib2005] S. M. T. Tabib. 2005. Some of grammatical structures are used in persian poems (برخی ساختارهای دستوری گونه‌ی شعری). *Persian Academy (Farhangestan)*, 1:65–78, February.

[Taghva et al.2003] K. Taghva, R. Beckley, and M. Sadeh. 2003. A list of farsi stopwords. Technical Report 2003–01, ISRI.

[Tousi1974] M. A. Adib Tousi. 1974. The affixes in persian language (وندهای فارسی). *Gohar*, 17:432–436, July.

[Wallach et al.2009] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. 2009. Evaluation methods for topic models. In *Proc. 26th An. Int. Conf. on Machine Learning (ICML'09)*, pages 1105–1112. ACM.

[Wallach et al.2010] H. Wallach, D. Mimno, and A. McCallum. 2010. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22 (NIPS'09)*, pages 1973–1981.