Exploring linguistically-rich patterns for question generation

Sérgio Curto
L^{2}F/INESC-ID Lisbon
sslc@l2f.inesc-id.pt

Ana Cristina Mendes
L^{2}F/INESC-ID Lisbon
IST, Tech. Univ. Lisbon
acbm@l2f.inesc-id.pt

Luísa Coheur
L^{2}F/INESC-ID Lisbon
IST, Tech. Univ. Lisbon
lcoheur@inesc-id.pt

Abstract

Linguistic patterns reflect the regularities of Natural Language and their applicability is acknowledged in several Natural Language Processing tasks. Particularly, in the task of Question Generation, many systems depend on patterns to generate questions from text. The approach we follow relies on patterns that convey lexical, syntactic and semantic information, automatically learned from large-scale corpora.

In this paper we discuss the impact of varying several parameters during pattern learning and matching in the Question Generation task. In particular, we introduce semantics (by means of named entities) in our lexico-syntactic patterns. We evaluate and compare the number and quality of the learned patterns and the matched text segments. Also, we detail the influence of the patterns in the generation of natural language questions.

1 Introduction

Natural Language (NL) is known for its variability and expressiveness. There are hundreds of ways to express an idea, to describe a fact. But language also comprises several regularities, or patterns, that denote the presence of certain information. For example, *Paris is located in France* is a common way to say that Paris is in France, indicated by the words *located in*.

The use of patterns is a widely accepted as an effective approach in the field of Natural Language Processing (NLP), in tasks like Question-Answering (QA) (Soubbotin, 2001; Ravichandran and Hovy, 2002) or Question Generation (QG) (Wyse and Piwek, 2009; Mendes et al., 2011).

Particularly, QG aims at generating questions from text and has became a vibrant line of research. Generating questions (and answers), on one hand, allows QA or Dialogue Systems to be easily ported to different domains, by quickly providing new questions to train the systems. On the other hand, it is useful for knowledge assessment-related tasks, by reducing the amount of time allocated for the creation of tests by teachers (a time consuming and tedious task if done manually), or by allowing the self evaluation of knowledge acquired by learners.

Most systems dedicated to QG are based on hand-crafted rules and rely on pattern matching to generate questions. For example, in (Chen et al., 2009), after the identification of key points, a situation model is built and question templates are used to generate questions. The Ceist system (Wyse and Piwek, 2009) uses syntactic patterns and the Tregex tool (Levy and Andrew, 2006) that receives a set of hand-crafted rules and matches the rules against parsed text, generating, in this way, questions (and answers). Kalady et al.(2010) bases the QG task in Up-keys (significant phrases in documents), parse tree manipulation and named entity recognition.

Our approach to QG also relies on linguistic patterns, defined as a sequence of symbols that convey lexical, syntactic and semantic information, reflecting and expressing a regularity of the language. The patterns associate a question to its answer and are automatically learned from a set of seeds, based on large-scale information corpora, shallow parsing and named entities recognition. The generation of questions uses the learned patterns, as questions are created from text segments found in free text after being matched against the patterns.

Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, pages 33–38,
Edinburgh, Scotland, UK, July 31, 2011. ©2011 Association for Computational Linguistics

This paper studies the impact on QG of varying linguistic parameters during pattern learning and matching. It is organized as follows: in Sec. 2 we introduce our pattern-based approach to QG; in Sec. 3 we show the experiments and discuss results; in Sec. 4 we conclude and point to future work.

2 Linguistically-Rich Patterns for Question Generation

The generation of questions involves two phases: a first offline phase – *pattern learning* – where patterns are learned from a set of seeds; and a second online phase – *pattern matching and question generation* – where the learned patterns are matched against a target document and the questions are generated. Next we describe these phases.

Pattern Learning Our approach to pattern learning is inspired by the work of Ravichandran and Hovy (2002), who propose a method to learn patterns based on a two-step technique: the first acquires patterns from the Web given a set of seeds and the second validates the patterns. Despite the similarities, ours and Ravichandran and Hovy's work have some differences: our patterns also contain syntactic and semantic information and are not validated. Moreover, our seeds are well formulated NL questions and their respective correct answers (instead of two entities), which allows to directly take advantage of the test sets already built and made available in evaluation campaigns for QA systems (like Text REtrieval Conference (TREC) or Cross Language Evaluation Forum (CLEF)).

We use a set of seeds, each composed by a NL question and its correct answer. We start by classifying each seed question into a semantic category, in order to discover the type of information these are seeking after: for example, the question *"Who painted the Birth of Venus ?"* asks for a person's name. Afterwards, we extract the phrase nodes of each seed question (excluding the Wh-phrase), enclose each in double quotes and submit them as a query to a search engine. For instance, given the seed *"Who painted the Birth of Venus ?"/Botticelli* and the syntactic structure of its question [$_{WHNP}$ Who] [$_{VBD}$ painted] [$_{NP}$ the Birth of Venus][1], we build the query: `"painted" "the Birth of Venus" "Botticelli"`.

We build patterns that associate the entities in the question to the answer from the top retrieved documents. From the sentence *The Birth of Venus was painted around 1486 by Botticelli*, retrieved as result to the above query, we learn the pattern "`NP VBD[was] VBN PP[around 1486]:`*`[Date]`*` IN:[by] NP{`*`ANSWER`*`}`"[2]. The syntactic labels without lexical information are related with the constituents of the question, while those with "{*`ANSWER`*}" mark the answer.

By creating queries with the inflected forms of the main verb of the question, we learn patterns where the surface form of the verb is different to that of the verb in the seed question (*e.g.*, "`NP{`*`ANSWER`*`} VBD[began] VBG NP`" is learned from the sentence *Botticelli began painting the Birth of Venus*). The patterns generated by verb inflection are INFLECTED; the others are STRONG patterns.

Our patterns convey linguistic information extracted from the sentences in the documents where all the constituents of the query exist. The pattern is built with the words, their syntactic and semantic classes, that constitute the segments where those constituents are found. For that, we perform syntactic analysis and named entity recognition in each sentence. In this paper, we address the impact of adding semantic information to the patterns, that is, the difference in having a pattern "`NP VBD[was]` *`VBN`* `PP[around 1486]:`*`[Date]`*` IN:[by] NP{`*`ANSWER`*`}`" with or without the named entity of type DATE, for instance.

Pattern Matching and Question Generation The match of the patterns against a given free text is done at the lexical, syntactic and semantic levels. We have implemented a (recursive) algorithm that explores the parsed tree of the text sentences in a top-down, left-to-right, depth-first search, unifying the text with the linguistic information in the pattern.

Also, we discard all matched segments in which the answer does not agree with the semantic category expected by the question.

The generation of questions from the matched text

[1] The Penn Treebank II Tags (Bies et al., 1995) are used.

[2] The patterns are more complex than the ones presented: they are linked to the seed question by indexes, mapping the position of each of its components into the question constituents.

segments is straightforward, since we keep track of the syntactic structure of the questions and the sentences on the origin of the patterns. There is a direct unification of all components of the text segment with the constituents of the pattern. In the INFLECTED patterns, the verb is inflected with the tense and person of the seed question and the auxiliary verb is also used.

3 Experiments

3.1 Experimental Setup

We used the 6 seeds shown in Table 1, chosen because the questions contain regular verbs and they focus on known entities – being so, it is probable that there will be several texts in the Web referring to them. However, understanding the characteristics of a pair that makes it a good seed is an important and pertinent question and a direction for future work.

GId: 1 **Syntactic Structure:** WHNP VBD NP **Semantic Category:** HUMAN:INDIVIDUAL
"Who wrote Hamlet?"/Shakespeare *"Who painted Guernica?"/Picasso* *"Who painted The Starry Night?"/Van Gogh*
GId: 2 **Syntactic Structure:** WHADVP VBD NP VBN **Semantic Category:** NUMERIC:DATE
"When was Hamlet written?"/1601 *"When was Guernica painted?"/1937* *"When was The Starry Night painted?"/1889*

Table 1: Seeds used in the experiments.

The syntactic analysis of the questions was done by the Berkeley Parser (Petrov and Klein, 2007) trained on the QuestionBank (Judge et al., 2006). For question classification, we used Li and Roth (2002) taxonomy and a machine learning-based classifier fed with features derived from a rule-based classifier (Silva et al., 2011).

For the learning of patterns we used the top 64 documents retrieved by Google and to recognize the named entities in the pattern we apply several strategies, namely: 1) the Stanford's Conditional Random-Field-based named entity recognizer (Finkel et al., 2005) to detect entities of type HUMAN; 2) regular expressions to detect NUMERIC and DATE type entities; 3) gazetteers to detect entities of type LOCATION.

For the generation of questions we used the top 16 documents retrieved by the Google for 9 personalities from several domains, like literature (*e.g.*, Jane Austen) and politics (*e.g.*, Adolf Hitler). We do not have influence on the content of the retrieved documents, nor perform any pre-processing (like text simplification or anaphora resolution). The Berkeley Parser (Petrov and Klein, 2007) was used to parse the sentences, trained with the Wall Street Journal.

3.2 Pattern Learning Results

A total of 272 patterns was learned, from which 212 are INFLECTED and the remaining are STRONG. On average, each seed led to 46 patterns.

Table 2 shows the number of learned patterns of types INFLECTED and STRONG according to each group of seed questions. It indicates the number of patterns in which at least one named entity was recognized (W) and the number of patterns which do not contain any named entity (WO). Three main results of the pattern learning phase are shown: 1) the number of learned INFLECTED patterns is much higher than the number of learned STRONG patterns: nearly 80% of the patterns are INFLECTED; 2) most of the patterns do not have named entities; and 3) the number of patterns learned from the questions of group 1 are nearly 70% of the total number of patterns.

	INFLECTED		STRONG		
GId	WO	W	WO	W	TOTAL
1	127	19	36	8	
	146		44		190
2	40	26	10	6	
	66		16		82
All	167	45	46	14	
	212		60		272

Table 2: Number of learned patterns.

The following are examples of patterns and the actual sentences from where they were learned:

– "`NP{ANSWER} VBZ NP`": an INFLECTED pattern learned from group 1, from the sentence *1601 William Shakespeare writes Hamlet in London.*, without named entities;

– "`NP VBD VBN IN[in] NP{ANSWER}`": a

STRONG pattern learned from group 2, from the sentence *(Guernica was painted in 1937.)*, without named entities;

– “NNP VBZ[is] NP[a tragedy] ,[,] VBN[believed] VBN IN[between] NP[1599]:*[NUMERIC_COUNT,NUMERIC_DATE]* CC[and] NP{ANSWER}”: an INFLECTED pattern learned from group 2, from the sentence *William Shakespeare’s Hamlet is a tragedy , believed written between 1599 and 1601*, with 1599 being recognized as named entity of type NUMERIC_COUNT and NUMERIC_DATE.

3.3 Pattern Matching and Question Generation Results

Regarding the number of text segments matched in the texts retrieved for the 9 personalities, Table 3 shows that, from the 272 learned patterns, only 30 (11%) were in fact effective (an effective pattern matches at least one text segment). The most effective patterns were those from group 2, as 12 from 82 (14.6%) matched at least one instance in the text.

GId	INFLECTED	STRONG	TOTAL
1	13	5	18
2	9	3 (2 W)	12
All	22	8	30

Table 3: Matched patterns.

Regarding the patterns with named entities, only those from group 2 matched instances in the texts. The pattern that matched the most instances was “NP{ANSWER} VBD NP”, learned from group 1.

In the evaluation of the questions, we use the guidelines of Chen et al. (2009), who classify questions as *plausible* – if they are grammatically correct and if they make sense regarding the text from where they were extracted – and *implausible* (otherwise).

However, we split plausible questions in three categories: 1) P$_a$ for plausible, anaphoric questions, *e.g.*, *When was she awarded the Nobel Peace Prize?*; 2) P$_c$ for plausible questions that need a context to be answered, *e.g.*, *When was the manuscript published?*; and 3) P$_p$, a plausible perfect question. If a question can be marked both as PL$_a$ and PL$_c$, we mark it as PL$_a$. Also, we split implausible questions in: 1) IP$_i$: for implausible questions due to incompleteness, *e.g.*, *When was Bob Marley invited?*; and 2) IP: for questions that make no sense, *e.g.*, *When was December 1926 Agatha identified?*.

A total of 447 questions was generated: 31 by STRONG patterns, 269 by INFLECTED patterns and 147 by both STRONG and INFLECTED patterns. We manually evaluated 100 questions, randomly selected. Results are in Table 4, shown according to the type (INFLECTED/STRONG) and presence of named entities (W/WO) in the pattern that generated them.

	P$_a$	P$_c$	P$_p$	IP$_i$	IP	Total
INFLECTED						57
WO	2	0	27	23	5	
STRONG						13
W	**1**	0	**1**	0	**1**	
WO	1	2	3	3	1	
INFL/STR						30
WO	0	0	9	18	3	
All	4	2	40	44	10	100

Table 4: Evaluation of the generated questions.

46 of the evaluated questions were considered plausible and, from these, 40 can be used without modifications. From the 54 implausible questions, 44 were due to lack of information in the question. 69% (9 in 13) of the questions originated in STRONG patterns were plausible. This value is smaller for questions generated by INFLECTED patterns: 50.8% (29 in 57). Questions that had in their origin both a STRONG and a INFLECTED pattern were mostly implausible, only 9 in 30 were plausible (30%). The presence of named entities led to an increase of questions of only 3 (2 plausible and 1 implausible).

3.4 Discussion

The results concerning the transition from lexico-syntactic to lexico-syntactic-semantic patterns were not conclusive. There were 59 patterns with named entities, but only 2 matched new text segments. Only 3 questions were generated from patterns with semantics. We think that this happened due to two reasons: 1) not all of the named entities in the patterns were detected; and 2) the patterns contained lexical information that did not allow a match with the text (*e.g.*, “NP{ANSWER} VBD[responded]

`PP[in 1937]:textit[Date] WHADVP[when] NP[he] VBD NP`" requires the words *responded*, *when* and *he*.)

From a small set of seeds, our approach learned patterns that were later used to generate 447 questions from previously unseen text. In a sample of 100 questions, 46% were judged as plausible. Two plausible questions are: *"Who had no real interest in the former German African colonies?"*, *"When was The Road to Resurgence published?"* and *"Who launched a massive naval and land campaign designed to seize New York?"*.

The presence of syntactic information (a difference between ours and Ravichandran and Hovy's work) allows to relax the patterns and to generate questions of various topics: *e.g.*, the questions *"Who invented the telegraph?"* and *"Who directed the Titanic?"* can be generated from matching the pattern "`NP VBD[was] VBN IN:[by] NP{`*`ANSWER`*`}`" with the sentences *The telegraph was invented by Samuel Morse* and *The Titanic was directed by James Cameron*, respectively.

4 Conclusions and Future Work

We presented an approach to generating questions based on linguistic patterns, automatically learned from the Web from a set of seeds. We addressed the impact of adding semantics to patterns in matching text segments and generating new NL questions.

We did not detect any improvement when adding semantics to the patterns, mostly because the patterns with named entities did not match too many text segments. Nevertheless, from a small set of 6 seeds, we generated 447 NL questions. From these, we evaluated 100 and 46% were considered correct at the lexical, syntactic and semantic levels.

In the future, we intend to pre-process the texts against which the patterns are matched and from which the questions are generated. Also, we are experimenting this approach in another language. We aim at using more complex questions as seeds, studying its influence on the generation of questions.

Acknowledgements

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through the project FALACOMIGO (ProjectoVII em co-promoção, QREN n 13449).

Ana Cristina Mendes is supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/43487/2008).

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert Macintyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project.

Wei Chen, Gregory Aist, , and Jack Mostow. 2009. Generating questions automatically from informational text. In *The 2nd Workshop on Question Generation*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. 43^{rd} Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370. ACL.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: creating a corpus of parse-annotated questions. In *ACL-44: Proc. 21^{st} Int. Conf. on Computational Linguistics and the 44^{th} Annual Meeting of the Association for Computational Linguistics*, pages 497–504. ACL.

Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das. 2010. Natural language question generation using syntax and keywords. In *The 3^{rd} Workshop on Question Generation*.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC 2006*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proc. 19^{th} Int. Conf. on Computational Linguistics*, pages 1–7. ACL.

Ana Cristina Mendes, Sérgio Curto, and Luísa Coheur. 2011. Bootstrapping multiple-choice tests with thementor. In *CICLing, 12^{th} International Conference on Intelligent Text Processing and Computational Linguistics*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proc. Main Conference*, pages 404–411. ACL.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *ACL '02: Proc. 40^{th} Annual Meeting on Association for Computational Linguistics*, pages 41–47. ACL.

João Silva, Luísa Coheur, Ana Mendes, and Andreas Wichert. 2011. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35:137–154.

M. M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. In *Proc. 10th Text REtrieval Conference (TREC)*, pages 293–302.

Brendan Wyse and Paul Piwek. 2009. Generating questions from openlearn study units. In *The 2nd Workshop on Question Generation*.