A novel approach for automatic extraction of semantic data about football transfer in sport news

Quang-Minh Nguyen
Department of Electronics and Computer Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam, and
Tuan-Dung Cao
Department of Software Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

Received 14 March 2015
Revised 14 March 2015
Accepted 5 April 2015

Abstract

Purpose – The purpose of this paper is to propose an automatic method to generate semantic annotations of football transfer in the news. The current automatic news integration systems on the Web are constantly faced with the challenge of diversity, heterogeneity of sources. The approaches for information representation and storage based on syntax have some certain limitations in news searching, sorting, organizing and linking it appropriately. The models of semantic representation are promising to be the key to solving these problems.

Design/methodology/approach – The approach of the author leverages Semantic Web technologies to improve the performance of detection of hidden annotations in the news. The paper proposes an automatic method to generate semantic annotations based on named entity recognition and rule-based information extraction. The authors have built a domain ontology and knowledge base integrated with the knowledge and information management (KIM) platform to implement the former task (named entity recognition). The semantic extraction rules are constructed based on defined language models and the developed ontology.

Findings – The proposed method is implemented as a part of the sport news semantic annotations-generating prototype BKAnnotation. This component is a part of the sport integration system based on Web Semantics BKSport. The semantic annotations generated are used for improving features of news searching – sorting – association. The experiments on the news data from SkySport (2014) channel showed positive results. The precisions achieved in both cases, with and without integration of the pronoun recognition method, are both over 80 per cent. In particular, the latter helps increase the recall value to around 10 per cent.

Originality/value – This is one of the initial proposals in automatic creation of semantic data about news, football news in particular and sport news in general. The combination of ontology, knowledge base and patterns of language model allows detection of not only entities with corresponding types but also semantic triples. At the same time, the authors propose a pronoun recognition method using extraction rules to improve the relation recognition process.

Keywords Ontology, Semantic Web, Named entity recognition, Natural language processing, Rule based extraction, Semantic annotation

Paper type Research paper

This paper is an extended version of Quang-Minh *et al.*, 2014, "Automatic creation of semantic data about football transfer in sport news", Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2014), ACM, Hanoi, pp. 356-364.

International Journal of Pervasive Computing and Communications
Vol. 11 No. 2, 2015
pp. 233-252

1742-7371
DOI 10.1108/IJPCC-03-2015-0018

1. Introduction

Over the years, the Web has become a news publishing channel approaching a large quantity of consumers and has the equal significance of television. Apart from accessing news Web sites that belong to a television channel or a news broadcaster, users can also read news from different sources thanks to the integration systems. The common characteristic of these systems is the news is constantly updated, which is sometimes overwhelming for users when looking for information in such an enormous data warehouse. As a result, users may encounter difficulty in reaching news matching their needs or personal preferences.

Accurate information searching, appropriate news organizing and classifying according to the concepts, improving navigation experience are the objectives which the developers of news systems are aiming to. However, these objectives are hindered by the model of representing and storing information based on structure and syntax. For example, news searching based on only keyword could cause interference and return false or incomplete news. News collected from multiple non-homogeneous sources can be duplicated or missing necessary association, while there are relations in their content.

Semantic Web (Berners-Lee *et al.*, 2001) is an extension of the current Web, allowing a computer to be able to "understand" the information on the Internet to provide better service for finding, integrating, reusing information and Web resources. Thus, the application of Semantic Webs in news synthesis, search and publication is a promising research issue. The BBC is one of the first media companies interested in the research on Semantic Web applications and development of news publishing systems, by building World Cup 2010 Web site under the dynamic semantic publishing architecture (Rayfield, 2012).

They have conducted research (Kobilarov *et al.*, 2009) on solution with cross-linking the content of Web sites from different domains such as music, news using ontology and DBPedia (Mendes *et al.*, 2012).

Smartweb System (Buitelaar *et al.*, 2006) is a multi-dialog system arising answers from the Semantic Web services. The above systems have achieved positive results, showing the advantages of Semantic Web application. To build such systems, it is necessary to obtain the metadata with rich information about the news and the semantic annotation is the key for their success. In a research paper review by Pellegrini (2012), he discussed challenges and achievements in utilizing semantic metadata in the news production process. The value of metadata in the content value chain including content acquisition, content editing and content distribution has been analyzed and clarified by the author.

Originally, studies focus on developing manual semantic annotation tools to achieve high accuracy. A well-known online annotation tool for blog and e-mail contents is ZEMANTA (2014). It supports person to insert the tags and the links through recommendations. OntoMat Annotizer (Handschuh *et al.*, 2001) is a manual annotation editor that runs on a Web browser. It has an ontology guidance and a fact browser, which allows people to expand the ontology, for example, add a new instance. Users can select parts of the texts and then drop them on the desired ontology classes, or choose a class and initialize its properties values. Another manual semantic annotation editor, PinPart (Markovski *et al.*, 2012) is a browser extension for content bookmarker. It provides the user with automatic categorization and semantic annotation of user's content. The semantic data is generated transparently to users, while they bookmark a Web page by invoking external application programming interfaces (APIs) from Zemanta, OpenCalais and AlchemyAPI. However, manual annotation is also an

expensive process. When the volume of documents becomes higher, it leads to the annotation acquisition bottleneck. Semi-automatic and automatic semantic annotation methods and tools such as PANKOW (Cimiano *et al.*, 2004), AeroDAML (Kogut and Holmes, 2001) and knowledge and information management platform (KIM) (Popov *et al.*, 2003) are proposed to overcome this challenge.

In our previous research (Quang-Minh *et al.*, 2012), we introduced a Semantic Web approach for developing BKSport, a system of sports information integration. We believe the functionality and friendliness of a sport news Web site or portal if there is an underlying semantic platform supported. The main idea is to collect sport news from heterogeneous sources over the Web and to describe them in a unified and explicit model as semantic annotation using ontology. Semantic annotation helps support for semantic searching and information visualization (Slimani, 2013). This is also the key to perform the association between the news based on their content. This representation model is also a prerequisite for creating management features – reusing content at a low cost for Web site managers and editors.

For automatic semantic annotation creation, we focused on the recognition of named entities in sport domain. Instead of detecting them in some general types such as other system did, we match them to instances of specialized concepts in our ontology. An algorithm was proposed to populate sport knowledge base and induce the semantic about the main topic of news.

In sports, transfer is an attractive news segment for the readers. The news about a player moving from one club to another club or contract signing between two clubs is posted by different sources of news. Unlike information about match results or other sports information, football transfer information contains different semantic branches. An appropriate method to extract these semantic will help enrich metadata of news in BKSport system. These semantic annotations of this particular field of information will be exploited by the functionalities of the news portal (e.g. transfer information synthesis interface, semantic search, related news listing).

This paper presents a method for generating semantic annotations about transfer in football news. Concepts and properties related to football transfer are designed and imported to BKSport ontology. We defined language models to capture popular semantic triples representing facts and events in football transfers. Based on these models, extraction rules are built using vocabulary from ontology. To improve the results of semi-automatic semantic annotation process, we extend rules with pronoun recognition pattern.

The rest of this paper is organized as follows: In Section 2, we briefly present some studies and system for excerpting semantic information. In Section 3, we describe the overall architecture of our process. In Section 4, we focus on methods for identifying transfer semantic relations based on language models. Also, in this section, we propose a pronoun recognition method using set of JAPE rules to integrate into transfer relation recognition method. The test results and evaluations will be presented in Section 5. Section 6 is the conclusion and orientation for our research in the future.

2. Related work

The process of attaching semantic models and natural language together is referred to semantic annotation. Information Extraction is becoming a central technology for bridging the gap between unstructured text and metadata expressed using ontology. In

recent years many efforts (Cimino and Barnett, 1993; Nguyen *et al.*, 2007; Dung and Kameyama, 2007; Cimiano *et al.*, 2004; Dill *et al.*, 2003; Tymoshenko and Giuliano, 2010) have been made to develop automatic and semi-automatic semantic annotation systems. However, none of these systems are designed for working on sport domain.

Pankow System (Pattern-based Annotation through Knowledge on the Web) (Cimiano *et al.*, 2004) exploited surface model and data redundancy on the Web to automatically classify entities from the text for a given ontology. The models are phrases such as: <Concept> <Instance> (e.g. forward Messi) and <Instance> is a <Concept> (e.g. Messi is a football player). The system builds models by identifying all the proper names in the text (using Part-of-Speech Tagger) and combines each private name with each of 58 concepts from their ontology into a hypothesis. Each hypothesis is then tested with the Web via Google queries and the number of views, which is the measure to assess the precision of model. Best performance of fully automated system is 24.9, and 62.09 per cent under manual operation.

SemTag (Dill *et al.*, 2003) is the semantic annotation component of Seeker platform, for performing large-scale annotation of Web pages. It works with respect to a shallow ontology called TAP, which includes a range of lexical and taxonomic information about popular items. After annotating all possible mentions of instances from the TAP ontology, SemTag performs taxonomy-based disambiguation algorithm. It uses a vector-space model to assign the correct concept or to determine that this mention does not correspond to a concept in ontology. The best precision value of SemTag is about 82 per cent, while the recall value is unknown.

Harrington and Clark (2008) described Asknet system, an information extraction system for building a large-scale Semantic Web data from unstructured text. Procedure for extracting information in Asknet is as follows. Firstly, syntax of sentences in the text will be analyzed with C & C syntax analyzer. A named entity recognition (NER) stage is performed using the C & C NER tagger. After the sentences have been analyzed, Asknet uses a semantic analysis engine-Boxer to generate the first order logical representation. The system obtains the overall precision of 79.1 per cent.

Sun and Han (2014) proposed an algorithm based on kernel tree to extract relations between two entities. They proposed a new kernel tree, called "feature-enriched tree kernel" (FTK), to overcome the raw performances or unclear problems in traditional syntax tree to capture the semantic relation better. Overall precision of the method is about 81.2 per cent.

Abacha and Zweigenbaum (2010) introduced an approach to extract relations between entities in the field of medicine using the language model. They used MetaMap to extract named entities in medicine such as drug name and patient name. To extract the desired relations, they designed a language model based on the selection of PubMed Central articles. Their tests achieved the precision of 74.21 per cent.

Tymoshenko and Giuliano (2010) have proposed an approach to extract semantic relations between noun phrases (nominals), based on the combination of semantic information provided by ResearchCyc to handle shallow parser. On the "Multi-way classification of semantic relations between pairs of nominals" task at SemEval- 2010, the method achieved the overall measure value F1 of 77.62 per cent.

Muthu Lakshmi and Uma (2010) developed a system of electronic learning (e-learning) based on the Semantic Web for the sports domain, using resource description framework (RDF), ontology and model Web language ontology. Ontology

engineering methodology was applied to improve quality of concepts and relationship in ontology. The authors also discussed the different types of sports and the communication between learner-instructor through e-learning.

3. Annotation process

Transfer in football is one of the news segments attracting much attention of the readers. Consequently, the football transfer news is updated regularly and fully on the Web. In this context, most of significant information is expressed in textual data, which is written in natural language that only humans can understand. For instance, it would be useful if from the news article:

> Torino have signed Serbian goalkeeper Vlada Avramov following his release from Cagliari. The 35-year-old was a free agent after leaving the Sardinian club, where he had been the first-choice keeper for three seasons, in June.

We can create a semantic representation of the event as a triple: "Torino-Club" "sign contract with" "Vlada Avramov". Combined with existing semantic annotations as "Torino-Club" is the club participating in the Serie A and Vlada Avramov is "goal keeper", this would help the system have a better comprehension of the content of the news article.

Our aim is to detect and to extract semantic information automatically from the news. After that, it is transformed to N-TRIPLES formalization and stored in the semantic data repository. Then, these metadata could be retrieved, exploited later by semantic engine for the implementation of functionalities in the BKSport news aggregation system as introduced by Quang-Minh *et al.* (2012). We decided to use Allegrograph to replace Jena in semantic engine and semantic repository for performance objectives, such as query optimization and data caching.

Overview about the proposed annotation process is shown in Figure 1. Firstly, the data from the sports Web sites will be collected and preprocessed for redundant data elimination. The filtered information is passed to recognizing named entities module to detect the appearance of football players, coaches, clubs, agents, etc. in the news. Finally, the detected instances are matched to a set of predefined rules to identify relations (properties) related to football transfer. If there is a match, the output will be mapped to a corresponding property in BKsport ontology to generate semantic triple.

Details of each step will be described in each section below.

3.1 BKSport ontology

As defined by Gruber (1993): "ontology is an explicit specification of a (shared) conceptualization". In fact, abstract ontology is often modeled with the following components: concepts (e.g. PERSON, LOCATION), entities of concepts (e.g. John is-instance-of PERSON, America is-instance-of LOCATION), characteristics of concepts and entities (e.g. PERSON has-age, LOCATION has- coordinates), relations between concepts (e.g. CITY is-subclass-of LOCATION), relations between entities (e.g. John has-mother Marry), etc.

The basic principles defined by Gruber to design and build ontology are as follows:

- *Clarity and objectivity*: The terms should be defined by natural language using ontology in a clear and objective way.
- *Integrity*: The definition must be complete and denote meaning of a specific term.

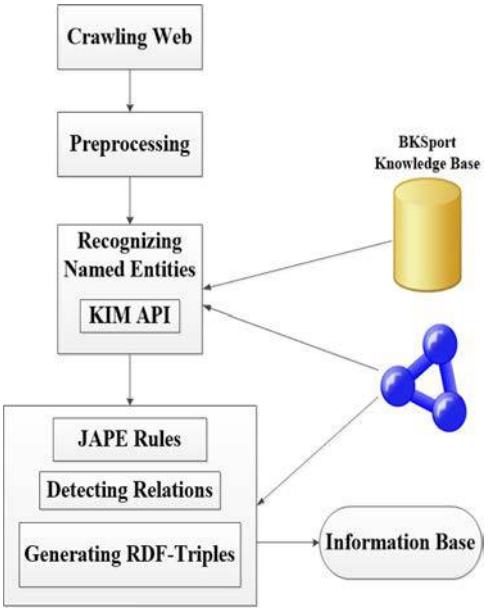

Figure 1.
Semantic relations extraction process

- *Consistency*: There is no conflict among the conclusions arising from argued knowledge and the semantics of the term.
- *Maximum one-way scalability*: It is not necessary to amend the existing terms when we add general or specific terms into ontology.
- *Minimum constraints*: Constraints in the model should be limited to as little as possible.

Certain studies on text analysis use ontology to improve the quality of entity recognition. OntoText has been developing KIM (Popov *et al.*, 2003), a system for NER for common sector domain that achieves high efficiency. KIM can be used to support automatic semantic annotations on the different types of content, with scalability for handling semi-structured documents or unstructured documents based on GATE (2014). The main idea of our approach is to reuse and to extend the capabilities of KIM for NER task for sports domain, with the help of a proper knowledge base. The results obtained are significant to the development of algorithms for the problem of extracting semantic triples from text. The core of KIM system is PROTON ontology. PROTON is developed within the framework of the SEKTproject (www.sektproject.com/), defining about 250 concepts and 100 properties to be able to provide almost all necessary concepts at a high level for semantic annotation, indexing and searching. However, PROTON ontology only defines general concepts and properties, not a specific domain. Thus, to identify transfer relations, we must build a sport ontology dedicated for sports field (BKSport ontology) and map it into PROTON ontology.

BKSport ontology plays a crucial role in transfer relations identification system. It describes the entities in the real world of sports fields and domains as well as characteristics and relationships between them. It has a full vocabulary set to describe

the basic information of sports news and transfer news. From the analysis of football transfer news, we identify key relations representing activity in football transfer. They are properties associating abstract concepts in our sport ontology, for example, SportPerson, SportTeam and between SportPerson and SportTeam. For example, SportPerson move-to SportTeam, SportTeam sign-with SportTeam, SportTeam concern-with SportPerson, Coach buy Defender. Figure 2 describes a partial vocabulary of BKSport ontology related to football transfer topic.

3.2 Step 1, 2: document crawling and preprocessing

The information will be obtained from sports Web sites, then be preprocessed to remove redundant information (e.g. advertising contents) and to keep the main content of the news. Title and related links are also retained because it could help the system identify and extract information more accurately.

3.3 Step 3: named entities recognizing

To understand semantics of the text, firstly, the system needs to understand semantics of the entities whose name appears in the text. The named entities in the sports domain include names of the players, coaches, clubs, stadiums, sports events, etc. For example: "Cordoba has completed the loan signing of Brazilian Winger Ryder Matos", the system needs to understand that Cordoba is a name of a football club and Ryder Matos is a name of a winger. To do this, there must be an identification step for named entity.

KIM is a platform that we reuse to identify named entities. KIM has been built to recognize entities in the public domain; it is not exclusive to a specific field. So to identify entities in the deeper and more detailed level in the sports domain, we need to add a set of concepts and properties in the ontology of KIM, and additional entities to the KIM knowledge base. In default ontology of KIM (PROTON ontology), named entities are represented at general level (e.g. person) not in detail (for example: winger, forward, etc).

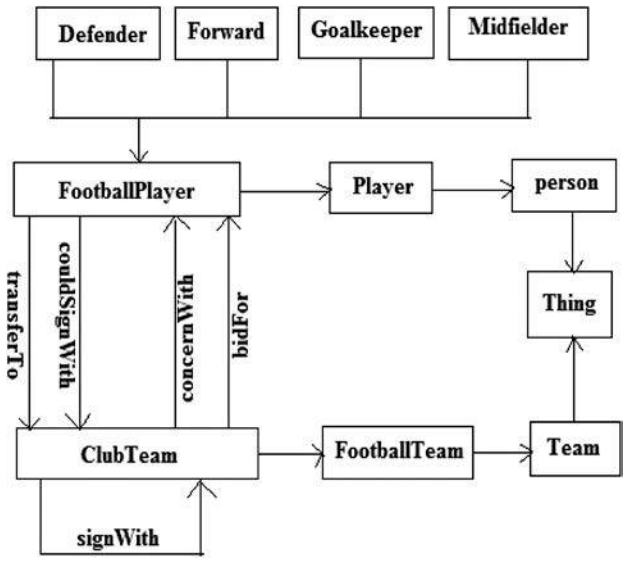

Figure 2.
A part of BKSport ontology

Therefore, we integrated BKSport ontology with PROTON, in the sense that more specialized concepts in the former will replace the abstract concept in the latter in recognition process. Thanks to the openness of the KIM platform, the integration could be done by mapping concepts between them. For example, classes of BKSport ontology such as coach, winger, forward and defender are understood as sub-classes of the person class. Figure 3 illustrates some classes mapped from BKsport ontology to PROTON ontology.

Ontology mapping does not assure the success of NER task without the additional knowledge base of football transfer news. To build this knowledge base, Web databases that contain information about football players, coaches, clubs and football agents in famous European football leagues are crawled and transformed to semantic annotation using BKSport ontology (Figure 4).

So far, we have supplemented the knowledge base of football players, coaches, stadiums, etc. of the Premier League, La Liga, Champions League; of tennis players from ATP rankings. The large number of instances in sport domain supplemented into knowledge base leads to a higher quantity of instances with the same names but different types. To resolve this ambiguity, we have defined a number of recognition rules to detect instances with name which does not appear in the knowledge base based on an observation that proper name in sport domain can come (immediately before or after) along with occupation as follows: <"occupation" + "private name"> or <"private name" + "occupation"> (e.g. winger Bale, Santiago Berbabeu stadium). Based on "occupation", the recognition rules detect the class which that named entity belongs to. For example, when an entity with its name appearing in the text, before which is "defender", it is identified as the name of a defender.

3.4 Step 4: extracting semantic relation

In this final step of the process, semantic properties appearing in the text will be captured by the language models that we have developed previously by a set of recognition rules. Each identified relation will be mapped to a corresponding relation in BKsport ontology to be expressed in the form of RDF triple or N-Triples. Details about the construction of identification rules will be presented in the next section (Figure 5).

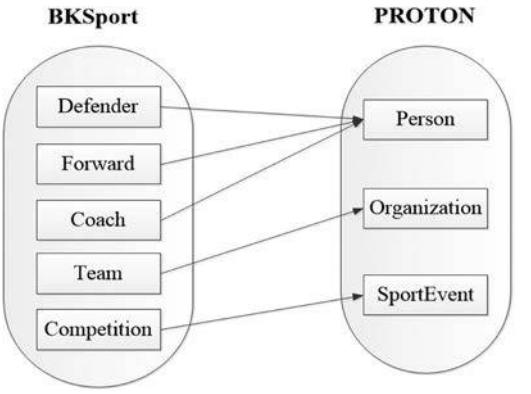

Figure 3.
Mapping from BKSport to PROTON

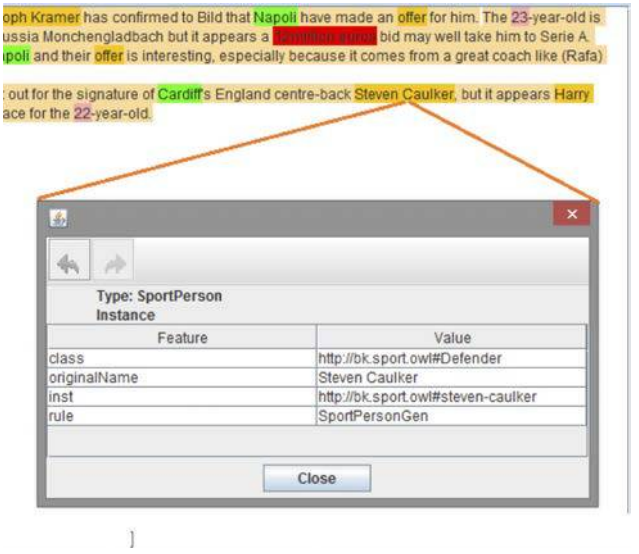

Figure 4.
Named entities recognition

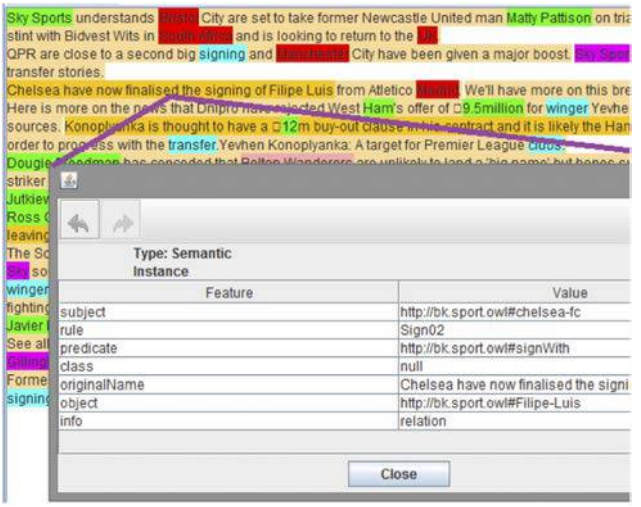

Figure 5.
Semantic relations recognition

For example: "Chelsea has now finalized the signing of defender Filipe Luis." The system is expected to understand that Chelsea have signed a contract with defender Filipe Luis, and map it to < bksport:signWith > relation, then express it in RDF:

```
<owl:Thing>
<rdf:subject rdf:resource="http://bk.sport.owl#chelsea-fc"/>
<rdf:predicate rdf:resource="http://bk.sport.owl#signWith"/>
<rdf:object rdf:resource="http://bk.sport.owl# filipe-luis "/>
</owl:Thing>
```

4. The method for extracting transfer semantic relation

4.1 Language models

The news is written in natural language, so the structure is very diverse and complex. Therefore, we can not build a set of models that can cover all the possible structures. But if the models are designed and constructed in an effective manner, it can bring a good result. Figure 6 presents our three main identification models.

As we are considering the football transfer domain, "Named Entity" will only often be person or team. The "phrasal verb" here is the phrase containing "verb" + "adverb" or "verb" + "preposition". The verbs characterize transfer relations, and the "tense" of the verb will determine that the relation belongs to one of the following three cases:

(1) the transfer took place;

(2) the transfer can happen in the near future; and

(3) the transfer was unsuccessful.

The "tense" of verb depends on the form of the verb or depends on words carrying meanings and standing before the verb. In this example: "Former Rangers goalkeeper Scott Gallacher has signed a two-year deal at Hearts", the verb "signed" shows that the transfer took place. Another example: "Barcelona forward Messi will make a new contract". The word "will" standing in front of the verb "make" to show that this case has not taken place, but may take place in the near future.

More specifically, in Figure 7, we represent the pattern for recognizing phrase verb as follows:

<ExtraVerb><Main Verb><Adverb/Preposition>

In which:

- "Extra Verb" are the words placed right before the main verb, indicates the action or transfer event falls into one of the following three cases: the event has not happened, but may happen in the new future, the event has happened and the event does not happen.
- "Main Verb" is the main verb of "phrasal verb".
- "Adverb/Preposition" is the following adverb and preposition modifying verb.

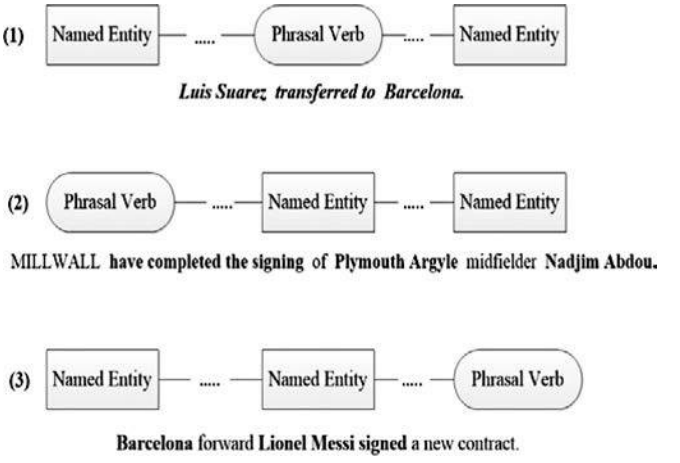

Figure 6.
Relations recognition patterns

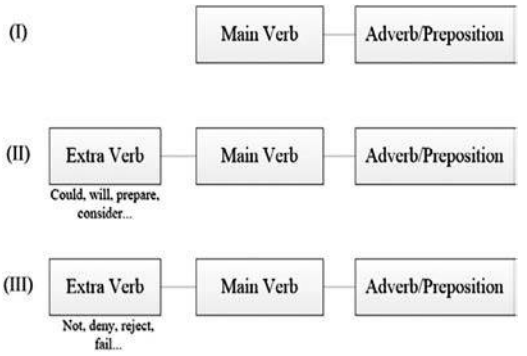

Figure 7.
Phrase verb modeling patterns

If there is no "extra verb" before "main verb", we assume that the event has happened (positive semantic)

If there is "extra verb" before "main verb", there are two cases: negative semantic and semantic representing possibilities:

- "extra verb" has the meaning of the near future, indicates the action or event may happen in future; e.g. "could", "prepare", "will", "consider"; and
- "extra verb" has negative meaning, indicates the action or event has not and will not happen; e.g. "not", "no", "don't", "fail", "reject".

4.2 Identification process of transfer relations

As described in the previous section, we use the recognition rules to recognize and extract transfer semantic relations. We chose JAPE (2012) as the language for rule representation because of its advantages. JAPE is a component of GATE, dedicated to identify defined entities by the rules, it is the language used to write the regular expressions via annotations.

Firstly, we need to split the text into sentences; each sentence contains certain content. The sentences usually begin and end with punctuation such as "." (Dot) "," (semicolon) or word (s) showing the start of a new content such as "while", "however", "but", so we can use the rules to do this easily. Then, each sentence will be matched with a list of rules.

We only consider named entities and phrasal verbs, so the unrelated words will be ignored. For phrasal verbs, because a verb can be in many different forms and many different words can express the same kind of semantic relation (e.g. "move to", "big moves", "transferred to" all express the relation "bksport:transferTo"), so while defining the rules, we also gather related verbs into sets of vocabulary. For example, a set of vocabulary that represents the signing is defined as follows:

```
Macro: SIGN
(
{Token.string==“sign”} | {Token.string==“signs”} | {Token.string=
=“signed”} | {Token.string==“signing”} | {Token.string==“signature”
}
)
```

Hereunder are two parts of two recognition rules, Sign01 and Transfer01:

```
Rule: Sign01
Priority: 80
(
   ({SportPerson}):p1
   ({Token.string!=".", Token.string!=";", !SportPerson})*
   (SIGN)
   ({Token.string!=".", Token.string!=";", Token.string!=","})*
   ({SportPerson}):p2
):sign
```

```
Rule: Transfer01
Priority: 70
(
   ({SportPerson}):p
   ({Token.string!=".", !SportPerson})*
   (TRANSFER)
   ({Token.string!=".", !SportTeam})*
   ({SportTeam}):t
):transfer
```

To recognize in these two cases: the transfer event will happen in the near future and the transfer event does not happen, we rely on the models II and III constructed in Section 4.1. Accordingly, apart from recognizing the main verb in news as presented, we also have to recognize the "extra verb". Corresponding to the two cases above, we create two "extra verb" vocabulary sets. The former contains the word (phrase) which represents the event will happen in near future:

```
Macro: COULD
({Token.string==“could”} | {Token.string==“will”} | {Token.string=
=“prepare”} | {Token.string==“consider”} | […])
```

The latter set contains a word (phrase) which represents the event does not happen:

```
Macro: NOT
({Token.string==“not”} | {Token.string==“deny”} | {Token.string==
“reject”} | {Token.string==“fail”} | […])
```

The following are two simple rules Sign02 and Sign03 for recognizing semantics which belongs to the two cases above:

```
Rule: CouldSign01
Priority: 90
(
   ({SportPerson}):p1
   ({Token.string!=".", Token.string!=";", !SportPerson})*
   (COULD)
   ({Token.string!=".", Token.string!=";", !SportPerson})*
   (SIGN)
   ({Token.string!=".", Token.string!=";", Token.string!=","})*
   ({SportPerson}):p2
):couldsign
```

```
Rule: NotSign01
Priority: 100
(
   ({SportPerson}):p1
   ({Token.string!=".", Token.string!=";", !SportPerson})*
   (NOT)
   ({Token.string!=".", Token.string!=";", !SportPerson})*
   (SIGN)
   ({Token.string!=".", Token.string!=";", Token.string!=","})*
   ({SportPerson}):p2
):notsign
```

If one text paragraph matches several rules, we will process to choose the most appropriate rule according to principles as follows:

- If the rules match one zone of document from point X, the one which matches the longest region will be chosen. For example, with two rules above (Sign01 and Transfer01) assume we have the text "Lionel Messi signed a contract with David Dein to move to Arsenal in the next season", rule Transfer01 would apply because it matches a longer region of text starting at the same point: "Lionel Messi signed a contract with David Dein to move to Arsenal". Meanwhile, rule Sign01 matches just "Lionel Messi signed a contract with David Dein".
- If the rules match one zone of document and have similar length, the rule with the highest priority will be chosen (we assigned to each rule a specified priority value, for example, with two rules above, rule Sign01's priority is 80 and rule Transfer01's priority is 70).
- If the rules have similar priority, the very first defined rule will be chosen.
- If all conditions above are similar, a rule will be chosen randomly.

Finally, the rules will map the identified relations to a corresponding relation in ontology to generate the RDF expression.

4.3 Pronouns annotation

In long text, to avoid repeating name of entities many times, pronouns are usually used to replace them. This causes direct difficulties in identifying of semantic relations because we must firstly identify named entities beforehand to identify the relations.

There are a number of studies on pronoun recognition. Qiu *et al.* (2004) carried out a study to describe an independent implementation, which was widely publicized by Resolution of Anaphorea Procedure developed by Lappin and Leass. It handles the third person pronouns, lexical anaphors, and recognizes redundant pronouns (pleonastic pronouns) in English language, obtaining the precision of 57.9 per cent (input format: MUC-6).

Liang and Wu (2004) proposed an anaphora resolution system which is based on the WordNet ontology and heuristic rules. The proposed system is capable of handling intra-sentential and inter-sentential anaphora in English text with appropriate handling for redundant pronouns. The system obtains overall success rate of 77 per cent.

Proposed method

Our method relies on the pattern based information extraction rules. They are only applicable in sports domain. Our set of rules which is built to denote pronouns, must comply with the following principles:

(1) Pronouns such as "he", "him", "I" and "me" represent SportPerson. Pronouns such as "they", "them", "we" and "us" represent SportTeam.

(2) Pronouns such as "I", "me", "we" and "us" appearing in indirect statements represent agents (SportPerson or SportTeam) which make that statement. There are two forms of indirect statement:

- Agent standing in front of indirect statement.
- Agent standing behind indirect statement.

(3) The pronouns representing named entities (SportPerson or SportTeam) appear in front of or near such pronoun (in case of indirect statement, the pronoun may represent entities behind it).

(4) After recognizing the pronouns, the rule will reset class field of the pronouns into class field of entity represented by it, to support the identification of transfer relations.

Besides, transfer news usually uses other special phrases to represent named entities. For example, use <'the' + number-year-old> to represent football players mentioned before. Considering this news:

> Inter Milan continues to work on new signings and reports in Italy claim there has been contact with Bundesliga side Hoffenheim regarding a deal for **Roberto Firmino**. *The 22-year-old* Brazilian attacking midfielder has previously been linked with the likes of Liverpool, and Hoffenheim reportedly want $7million (£5.5m) for him.

In this news, phrase "The 22-year-old" is used to replace Roberto Firmino, as shown in Figure 8.

5. Experiment and evaluation

Our work is among the first efforts on extracting semantic relations in football transfer. There is no standard dataset for us to test the proposed method and make comparisons

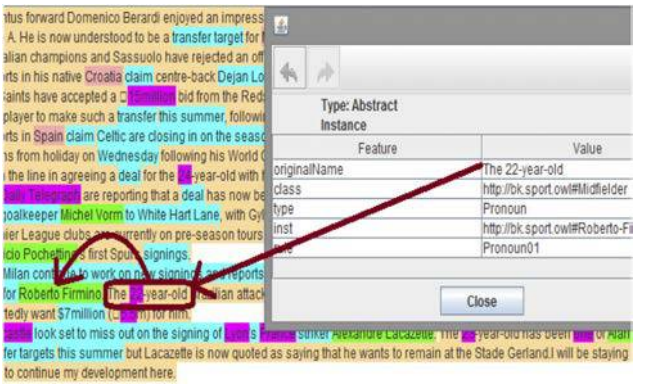

Figure 8.
Example results of pronouns recognition

with previous results. To evaluate our approach, we setup the experimentation as follows:

- News was crawled from Sky Sports (2014) page and stored in dataset to be processed separately.
- We conducted the semantic annotation for every news created by humans, after that these detected relations are marked.
- Program implementing proposed method was executed on every news in dataset. The relations recognized by the system will be compared with those marked earlier and we measured the right and the wrong ones.
- Total number of relations recognized and the total number of right relations identified by the system are collected to calculate precision and recall value.

We built a dataset that contains 237 football transfer news from SkySport Web site. We do the annotation task manually and detect 264 triples on football transfer. We carried out the experimentation in two scenarios:

(1) Do not use pronoun recognition rules.

(2) Use pronoun recognition rules.

In this news paragraph: "Torino have signed Serbian goalkeeper Vlada Avramov following his release from Cagliari. The 35-year-old was a free agent after leaving the Sardinian club", we can see that the news has two transfer semantic relations: the first one is "signWith" semantic relation between Torino club and goalkeeper Vlada Avramov; the second is "leave" semantic relation (goalkeeper Vlada Avramov leaves Sardinian club). However, in case the pronoun is not used, the system only identifies the first semantic relation because it uses "The 35-year-old" to replace the goalkeeper Vlada Avramov.

To evaluate the two methods, we use two criteria: recall and precision. Recall is defined as the rate of semantic relation that the system identifies correctly over the total relations which are identified manually. Precision is defined as the ratio of accurately identified semantic relations over the total relations identified by the system:

$$Precision\,(P) = \frac{Relevant\ retrieved\ relations\ (RRR)}{Total\ regconized\ relations\ (TRG)}$$

$$Recall\,(R) = \frac{Relevant\ retrieved\ relations\ (RRR)}{Total\ relevant\ relations\ (TRL)}$$

Table I shows the initial experimental results obtained from the first implementation of the method (Quang-Minh *et al.*, 2014). Some semantic triples are not identified due to complex structure, as the corresponding relations have ambiguous meanings

Table I. Initial performance of semantic relations recognition

	TRL	TRG	RRR	P%	R%
Case(1)	264	167	134	80.2	50.8
Case(2)	264	195	158	81.0	59.8

(Figure 11). For example, "Queens Park Rangers boss Harry Redknapp is eyeing a reunion with former Tottenham star Rafael van der Vaart".

Some cases were misidentified due to the following reasons: in sentences, there are a number of similarly named entities at once, and the system could not recognize the main entity of the relation; context information (describing what has not happened and negative events) is not included in the keywords, but lies in the meaning of the sentence. For example, the message: "The odds on Antoine Griezmann joining Monaco have shortened again". The system identifies as <Antoine Griezmann> <transferTo> <Monaco>, but in fact this event has not happened, as shown in Figure 10 (Figures 9-11).

Analyzing results of the first experiment, we find out that the recall is not high because of the complex structure of sentences and the quality of vocabulary set which is used for recognizing phrase verb. We made small improvements by reviewing relations in the ontology and adding synonyms and variant forms of a verb into the vocabulary. In addition, a step of sentence preprocessing has been carried out to transform the possessive case to the standard form, for example *<Named Entity>'s signature* is transformed to *the signature of <Named Entity>*. Thanks for this step, more semantics about football transfer can be captured by the existing recognition rules.

Figure 9.
Example of correct recognized semantics

Figure 10.
Example of incorrect recognized semantics

Figure 11.
Example of unrecognized semantics

The Table II shows the experimental results obtained from the above efforts. A greater coverage of about 5 per cent is observed while the precision does not change much.

In comparison to other works related to extracting semantic information in general domain as well as specific domain such as: PANKOW (Cimiano *et al.*, 2004) (maximum precision 69 per cent), KIM (precision 86 per cent, Recall 82 per cent), SemTag (precision 82 per cent) and Asknet system (Harrington and Clark, 2008) (total precision 79.1 per cent), approach utilizing kernel tree by Sun and Han (precision 81.2 per cent), method utilizing language model to extract relations between entities in medicine by Abacha and Zweigenbaum (2010) (precision 74.21 per cent), the preliminary results of this research are promising. Moreover, it is convinced that when we apply automatic annotation methods that are invented to work on general domain on a specific domain such as sport, the results are modest. As KIM or ASKNET platform can not recognize named entity of any famous football player in their professional context, it is nearly impossible to generate automatically annotations about the transfers they are concerning.

Figure 12 shows the output of our prototype. Extracted semantic triples are exported in N-triples format and it is quite simple to transform them to other formalization such as RDF or OWL.

Table II. Improved performance of semantic relations recognition

	TRL	TRG	RRR	P%	R%
Case(1)	264	180	145	80.5	54.9
Case(2)	264	213	173	81.2	65.5

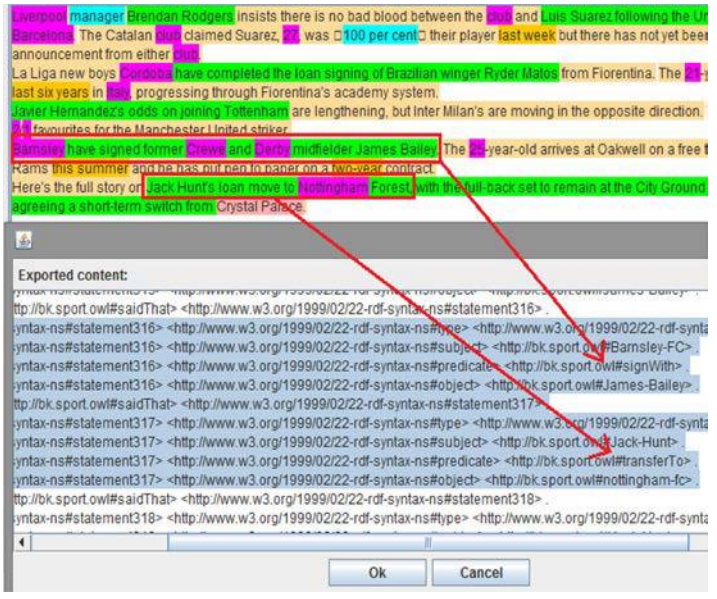

Figure 12. Semantic triples extracted as results

6. Conclusion

This work focuses on the presentation of a method to extract semantic relations in football transfers news using language models. We have reused KIM platform in the NER task because of two reasons: its openness and its impressive coverage when applied to general domain text in comparison to other systems. Primitive vocabularies for representing football transfer information are designed and incorporated into BKSport Ontology. The integration of our ontology with PROTON ontology has guided KIM to identify named entities as instances of concepts at more specialized levels in football domain. The language models are built on the recognition rules to capture the semantic relations. To improve the recall, we further propose an entity co-reference resolution method that relies on pronouns identification. The experiment on data set constructed from news taken from Sky Sports (2014), achieved relatively good precision.

From the observation of lexical and syntactic regularity of the football transfers domain, we can argue that rule creation is not time consuming for capturing the frequent semantics relation. We are presently working along several lines of development. On the one hand, we aim to evaluate our approach thoroughly with long-term experiments as well as continue enriching knowledge base and improving the rules of recognition to obtain better results.

Currently, new recognition samples are aiming to detect explicit semantics, whereas implicit semantics are still widely used in the journal writing style. Combining the deduction ability on recognizable semantic data and knowledge base could be a potential direction. Let us take a news article "Good news for Everton fans as the Mail is reporting that Romelu Lukaku could be close to a permanent move to Goodison Park", for example. If it is provided in the knowledge base that Goodison Park is the home of Everton, it is entirely possible to deduce that the transfer to this club belongs to Lukaku.

On the other hand, we are also interested in the hybrid rule-based approach for semantic annotation, which automates the creation of extraction rules with machine learning algorithms. There are efforts in this direction, especially for the specific domain application such as biology (Bannour *et al.*, 2013) and medical (Embarek and Ferret, 2008). Finally, with the benefits of the Linked Data paradigm compared to the conventional data integration approaches (Auer, 2011), we hope to build a linked open data resource in football and sport fields.

References

Abacha, A.B. and Zweigenbaum, P. (2010), "Automatic extraction of semantic relations between medical entities: a rule based approach", *Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)*, *Hinxton*, 25-26 October.

Auer, S. (2011), "Creating knowledge out of interlinked data", *Proceedings of WIMS'11*, *Sogndal*, *25-27 May*, pp. 1-8.

Bannour, S., Audibert, L. and Soldano, H. (2013), "Ontology-based semantic annotation: an automatic hybrid rule-based method", *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, August, pp. 139-143.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American Magazine*, 17 May.

Buitelaar, P., Cimiano, P. and Weber, N. (2006), "Ontology learning and population in smartweb", *Philips Symposium on Intelligent Algorithms (SOIA)*, *Eindhoven*.

Cimiano, P., Handschuh, S. and Staab, S. (2004), "Towards the self-annotating web", *Proceedings of the 13th International World Wide Web Conference*, *New York, NY, 17-20 May*, pp. 462-471.

Cimino, J. and Barnett, G. (1993), "Automatic knowledge acquisition from medline", *Methods of Information in Medicine*, Vol. 32 No. 2, pp. 120-130.

Dill, S., Gibson, N., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A. and Zien, J.Y. (2003), "Semtag and seeker: bootstrapping the semantic web via automated semantic annotation", *Twelfth International World Wide Web Conference*, Budapest, pp. 178-186.

Dung, T. and Kameyama, W.A. (2007), "Proposal of ontology-based health care information extraction system: VnHIES", *Research, Innovation and Vision for the Future, 2007 IEEE International Conference*, *Hanoi*, pp. 1-7.

Embarek, M. and Ferret, O. (2008), "Learning patterns for building resources about semantic relations in the medical domain", *Proceedings of the Language Resources and Evaluation Conference*, *Marrakech*.

GATE (2014), available at: https://gate.ac.uk/

Gruber, T.R. (1993), "A translation approach to portable ontology specifications", *Knowledge Acquisition*, Vol. 5 No. 2, pp. 199-220.

Handschuh, S., Stabb, S. and Maedche, A. (2001), "Cream – creating relational metadata with a component based, ontology-driven annotation framework", *Proceedings of K-Cap 2001*, *Victoria*, ACM Press, pp. 76-83.

Harrington, B. and Clark, S. (2008), "Asknet: creating and evaluating large scale integrated semantic networks", *International Journal on Semantic Computing*, Vol. 2 No. 3, pp. 343-364.

JAPE (2012), available at: http://gate.ac.uk/sale/tao/splitch8.html

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C. and Lee, R. (2009), "Media meets semantic web – how the BBC uses DBpedia and linked data to make connections", *ESWC 2009 Heraklion Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, *Heraklion*, 31 May-4 June, pp. 723-737.

Kogut, P. and Holmes, W. (2001), "AeroDAML: applying information extraction to generate DAML annotations from web pages", *First International Conference on Knowledge Capture (K-CAP 2001), Workshop on Knowledge Markup and Semantic Annotation*, Victoria.

Liang, T. and Wu, D.S. (2004), "Automatic pronominal anaphora resolution in English texts", *International Journal of Computaional Linguistics and Chinese Language Processing*, Vol. 9 No. 1, pp. 1-20.

Markovski, A., Jovanovik, M. and Trajanov, D. (2012), "Web extensions for semantic data creation", *9th International Conference for Informatics and Information Technology – CIIT 2012*, *Molika, Bitola*, pp. 125-128.

Mendes, P., Jakob, M. and Bizer, C. (2012), "DBpedia: a multilingual cross-domain knowledge base", *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, *Istanbul*, pp. 1813-1817.

Muthu Lakshmi, S. and Uma, G.V. (2010), "Semantic web based e-learning system for sports domain", *International Journal of Computer Applications*, Vol. 8 No. 14.

Nguyen, D.P.T., Matsuo, Y. and Ishizuka, M. (2007), "Exploiting syntactic and semantic information for relation extraction from wikipedia", *IJCAI Workshop on Text-Mining & Link-Analysis*, *Hyderabad*.

Pellegrini, T. (2012), "Semantic metadata in the news production process – achievements and challenges", *Proceeding from the 16th International Academic MindTrek Conference*, *Tampere*, pp. 125-133.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003), "KIM – semantic annotation platform", *2nd International Semantic Web Conference*, FL, pp. 834-849.

Qiu, L., Kan, M.Y. and Chua, T.S. (2004), "A public reference implementation of the RAP anaphora resolution algorithm", *Proceedings of the Language Resources and Evaluation Conference 2004*, *Lisbon*, pp. 291-294.

Quang-Minh, N., Tuan-Dung, C., Hoang-Cong, N. and Hagino, T. (2012), "Towards efficient sport data integration through semantic annotation", *Proceeding of The Fourth International Conference on Knowledge and Systems Engineering*, *Da Nang Viet Nam*, pp. 99-106.

Quang-Minh, N., Tuan-Dung, C. and Tam-Thanh, N. (2014), "Automatic creation of semantic data about football transfer in sport news", *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services*, *Hanoi*, pp. 356-364.

Rayfield, J. (2012), "Sports refresh: dynamic semantic publishing", available at: www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html (accessed 20 April 2012).

Sky Sports (2014), available at: www1.skysports.com/transfer-centre/

Slimani, T. (2013), "Semantic annotation: the mainstay of semantic web", *International Journal of Computer Applications Technology and Research*, Vol. 2 No. 6, pp. 763-770.

Sun, L. and Han, X. (2014), "A feature-enriched tree kernel for relation extraction", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, *Baltimore, MD*, 23-25 June, pp. 61-67.

Tymoshenko, K. and Giuliano, C. (2010), "FBK-IRST: semantic relation extraction using Cyc", *Proceedings of the 5th International Workshop on Semantic Evaluation*, *Uppsala*, pp. 214-217.

ZEMANTA (2014), available at: http://developer.zemanta.com/

Further reading

Chen, C.M. and Chen, L.H. (2014), "A novel approach for semantic event extraction from sports webcast text", *Multimedia Tools and Applications*, Vol. 71 No. 3, pp. 1937-1952.

Lee, C., Khoo, C. and Na, J. (2004), "Automatic identification of treatment relations for medical Ontology learning: an exploratory study", *Proceedings of the Eighth International ISKO Conference*, *London*, pp. 245-250.

Corresponding author

Quang-Minh Nguyen can be contacted at: minh.nguyenquang@hust.edu.vn