Recovery of Temporal Expressions From Text: The RISO-TT Approach

Adriano Araújo Santos
Programa de Pós-Graduação da Universidade Federal de Campina Grande - UFCG
Faculdade de Ciências Sociais Aplicadas - FACISA
Universidade Estadual da Paraíba - UEPB, Paraíba, Brazil
adriano@copin.ufcg.edu.br

Ulrich Schiel
Departamento de Sistemas e Computação – Universidade Federal de Campina Grande - UFCG
Campina Grande, Paraíba, Brazil
ulrich@computacao.ufcg.edu.br

***Abstract*— The necessity of managing the large amount of digital documents existing nowadays, associated to the human inability to analyze all this information in a fast manner, led to a growth of research in the area of development of systems for automation of the information management process. Nevertheless, this is not a trivial task. Most of the available documents do not have a standardized structure, hindering the development of computational schemes that can automate the analysis of information, thus requiring jobs of information conversion from natural language to structured information. For such, syntactic, temporal and spatial pattern recognition tasks are needed. Concerning the present study, the main objective is to create an advanced temporal pattern recognition mechanism. We created a rules dictionary of temporal patterns, developing a module with an extendable and flexible architecture for retrieval and marking. This module, called RISO-TT, implements this pattern recognition mechanism and is part of the RISO project (Retrieval of Information with Semantics of Contexts). Two experiments were carried out in order to evaluate the efficiency of the approach. The first one was intended to verify the extendibility and flexibility of the RISO-TT architecture and the second one analyses the efficiency of the approach, based on a comparison between the developed module and two consolidated tools in the academic community (Heidelime and SUTime). RISO-TT outperformed the rivals in the temporal expression marking process, which was proved through statistical tests.**

Keywords-temporal expressions extractor; temporal pattern recognition; natural language processing

I. INTRODUCTION

The large amount of digital documents existing nowadays, resulting from the unconstrained access and freedom to publish given by the Web to people [1] defeated the human capability to analyze all existing information. So, there is an increasingly need for automating the access, research, and management of information in order to generate valuable sources of knowledge [9].

Computers, in turn, can process structured or semi-structured information. Nevertheless, since most of the available information are non-structured [9], the present challenge is to allow computers to process information in natural language by converting this natural language to structured information, thus allowing a higher level of automation of computational processes.

So, Natural Language Processing (NLP) emerges as a possible solution for this challenge, because it is characterized as a set of computational techniques for analysis of texts in one or more linguistic levels, with the purpose of simulating the human linguistic processing. Among such techniques, there is the Recognition of Mentioned Entities (RME) [4], which aims to locate and classify atomic elements of a text, according to a pre-defined set of categories.

Information Extraction (IE) is the task of retrieving information from large volumes of documents or texts, structured or free [17]. For Zambenedetti [17], a well-developed information extraction technology allows the rapid development of extraction systems for new tasks that would have the same performance level of tasks performed by humans, a level not reached yet.

This paper addresses the process of extracting temporal expressions from texts, which is an activity that has become a significant research and development field in Computer Science, motivated by the large number of applications that explore temporal information extracted from texts. As examples of such applications we may cite Geographic Information Systems, automatic question, and answer applications and text summarization systems. With the use of temporal expression extraction techniques, applications perform tasks in a higher automation level [12].

There are several approaches for the recognition of temporal expressions in texts. Saquete [12] enumerates as main approaches: a) rules-based; b) Machine Learning and c) Combining rules, and machine learning. Regardless of the adopted approach, the output is a scheme of standardized temporal annotations. The schemes TIDES 2005 (Translingual Information Detection, Extraction, and Summarization) [11] and TimeML [14] are the most adopted.

Considering some limitations of the evaluated existing tools, we developed a new temporal expressions extractor, Retrieval of Semantic Information from Textual Objects Temporal Tagger (RISO-TT), as part of a project called RISO (Retrieval of Information with Semantics of Contexts) [18].

We carried out an experiment to prove the extensibility and flexibility of the system, as well as to check whether RISO-TT presents some competitive advantage and brings any contribution for this research area.

 ISBN: 978-1-61208-293-6

In the following section other approaches of temporal expressions extraction are presented. Section III introduces the RISO-TT approach to temporal expressions extraction and after that, the results of the proposal are analyzed on section IV Verification and Validation. Contributions and future work are discussed in the final section V Conclusions.

II. RELATED WORKS

Developed at the Center for Computational Language and Education Research, University of Colorado, the ATEL (Automatic Temporal Expression Labeler) [5] adopts a statistical approach to detect temporal expression in English and Chinese languages.

The system used a training database made available by TERN (Time Expression Recognition and Normalization) with a set of temporal terms and, for each sentence found in a processed document, the term is marked with tags.

ITC-IRST, Centro per la Ricerca Scientifica e Tecnológica, Povo, Italy, has developed Chronos System [10]. It is a rules-based approach, separating the identification of temporal expressions into recognition (detection) and interpretation of values (normalization). Chronos is based on linguistic analysis (tokenization, tagging and pattern recognition).

Another system, TempEx, has been eveloped by MITRE Corporation with a Perl application for recognition and interpretation of temporal expressions according to the TIMEX2 2001 specifications, TempEx is characterized as one of the firsts of this kind [8].

TempEx is able to recognize absolute times (E.g., March 15th, 2013) and relative ones (e.g., born after World War II), and the computation of the normalization is based on the publication date of the document, which means that the algorithm uses meta-information from the very document to compute the normalization of relative times [8].

Developed by the University of Georgetown, GUTime [16] is an extension of the TempEx tagger [8], recognizing and normalizing temporal expressions in TIMEX3 standard.

An important feature of this system is that it enables shifting temporal expressions, causing computations to be performed with basis on an input date [8]. GUTime has incorporated a set of ACE TIMEX2 expressions, including duration, a variety of temporal modifiers and European date formats [16].

The DANTE (Detection and Normalization of Temporal Expressions) system [9] has a modular architecture which consists, basically, of two modules: recognition and interpretation.

The temporal expression recognition module was developed by using of the JAPE grammar [Cunningham et al. 1999] which consists of a set of <condition, action> rules

The interpretation module scans sentence by sentence a document, searching for patterns that match a pre-defined one (knowledge base).

TERSEO (Temporal Expression Recognition System applied to Event Ordering) was developed by the Research Group on Natural Language Processing and Information Systems, University of Alicante. The system generates annotations in TIMEX2 standard.

At first, according to Saquete [12], TERSEO was developed as a knowledge base system, intended for automatic recognition and normalization of temporal expressions in Spanish texts. It uses the translation of the temporal expressions to temporal models already defined in the first version to obtain, automatically, the temporal expressions from other languages [12].

TIPSem [7] deals with six different tasks related to the treatment of multilingual temporal information proposed by TempEval-2, (Evaluating Events, Time Expressions, and Temporal Relations). These tasks are classified as A, B, C, D, E and F, where the task A consists in defining temporal extensions, task B consists in classifying the events defined by TimeML (Markup Language for Temporal and Event Expressions) and the remaining task are related to categorization of different temporal links.

Heideltime [13] is a rule-based system based intended to extract and normalize temporal expressions in several languages. It uses the TIMEX3 annotation standard, and there are, presently, versions in English and German languages. The marking of temporal expressions in HeidelTime depends on the domain where the documents are inserted, such as news, reports, colloquial, or science (e.g., biomedical studies).

The SUTime is a library for recognizing and normalizing temporal expressions developed by Stanford University. It is a system developed in Java and based on deterministic rules designed to be extensible.

In its development was used TokensRegex framework, a generic framework for defining patterns on the text and mapping of semantic objects and makes use of regular expressions for the recognition of temporal expressions [3].

For the analyzed tools, we also observed that most of the temporal expression extraction tools are neither flexible nor extensible, and the recognition of compound temporal expressions or not allowed.

III. RISO TEMPORAL TAGGER (RISO-TT)

RISO-TT is the temporal expression extractor of the RISO Project (acronym, in Portuguese, of Semantic Information Retrieval from Textual Objects). It differs from the other temporal extraction tools for considering more complex signs and grammatical associations in the process of identifying temporal expressions.

The complex temporal expressions considered result from grammatical associations, which determine time intervals and not just temporal tokens. Compound temporal expressions are more accurate because they allow the specific understanding of the time expressed in the text.

A compound temporal expression is a structure formed by closed intervals (e.g., *from the beginning of January 10th to July 20th*), or semi-open intervals (*since 1968*). Also grammatical associations formed by prepositions, adverbs, numbers, and temporal tokens are considered, such as *in December*, or *every day*, and several other relation of terms in a semantic temporal expression.

 ISBN: 978-1-61208-293-6

To exemplify a compound temporal expression and to show why is it necessary to be identified, consider the expression "*from December 10, 2011 to December 10, 2012*". This sentence refers to a specific period which can be represented by 12/10/2011 < X < 12/10/2012, where X is the temporal variable referred by the expression.

RISO-TT does not depend on fixed standards and third-party software in its architecture. Both cases can cause problems in the future, since standards evolve, and a considerable architectural change can lead to serious development problems.

A. Architecture

RISO-TT is a rules-based system and since it was conceived to become extensible and flexible it uses a configuration file which determines the connection from the internal logic to the rules-base. To insert a new standard (or rule) in the rules base, the new standard is simply added to the configuration files. For this change to be realized, it just needs that RISO-TT be run again with a document that contains, in its content, the corresponding expressions. Figure 1 presents the RISO-TT architecture.

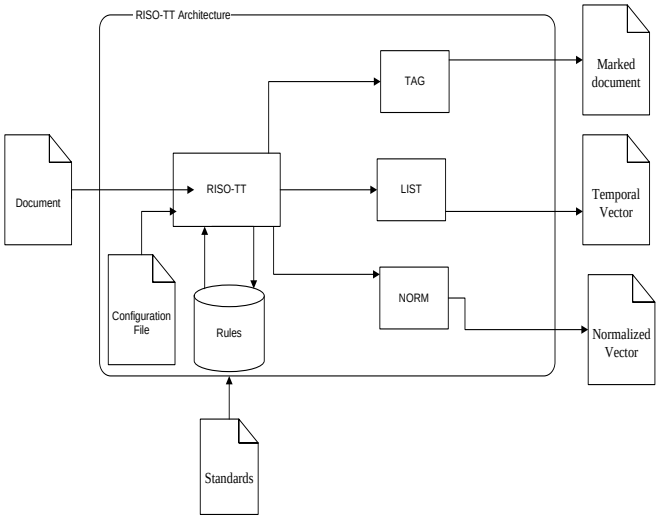

Figure 1. RISO-TT architecture.

A standard is a sequence of (temporal or grammatical) semantically associated terms, which can assign a value to a temporal expression. Typical standards are: preposition, adverbs, seasons of the year, dates, hours, and regular expressions.

A rule is an ordered sequence of standards (temporal and/or nominal) that characterizes the formation of temporal expressions. A rule takes into consideration the position of terms that form an expression. For example, the rule *Day Month Year* is different from the rule *Month Day Year*.

With the use of rules, the temporal standards were extended in such a way that complex structures among the grammatical relations and classical temporal expressions can be recognized as a single expression. With this, expressions like "*from December 10, 2011 to December 10, 2012*" are classified as a single temporal expression, and not as two independent temporal tokens.

B. Processed Outputs

The processing of a document in RISO-TT generates three documents:

- Marked document (TAG): the document given as input generates a document marked with the RISOTime tags and type attributes (e.g., <RISOTime type=Pre-EBT>On September 1, 1939</RISOTime>). The value assigned to the type attribute is the name of the rule of which the expression found is part.
- Temporal Vector (LIST): a document with a list of the temporal expressions found in the document (e.g., EBT-N -> from 499 to 493 BC) is created.
- Normalized Vector (NORM): Another document with a list of the temporal expressions found in the input document and their normalized values (e.g., On September 1, 1939 <--> 1-09-1939).

C. Example of Document Processing in RISO-TT

The following sentence is part of the document "16_SpanishCivilWar" from WikiWars:

... On 7 March, the Nationalists launched the Aragon Offensive. By 14 April they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace in May, but Franco demanded unconditional surrender; the war raged on. In July, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, from 24 July until 26 November.

We can find many types of temporal expressions in these sentences and simple expressions such as *On 7 March* are found by common Temporal Expression Extractors. But some types of temporal expression are more complex (e.g., *from 24 July until 26 November*). Theses sentences we called of compound temporal expressions.

The RISO-TT finds simple and compound temporal expressions and, if an expression is not detected, a corresponding rule identifying this kind of expressions can be added to the rules base.

The example text above generates the following marked document to by the function TAG:

 ISBN: 978-1-61208-293-6

<RISOTime type=Pre-EBT>On 7 March</RISOTime>, the Nationalists launched the Aragon Offensive. <RISOTime type=Pre-EBT>By 14 April</RISOTime>, they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace <RISOTime type=Pre-EBT>in May </RISOTime>, but Franco demanded unconditional surrender; the war raged on. <RISOTime type=Pre-EBT>In July</RISOTime>, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, <RISOTime type=I>from 24 July until 26 November</RISOTime>.

where 'type=Pre-EBT' means the association of a *Preposition* and a *Basic Temporal Structure*. "type=I**"** is the Intervals rule.

The function LIST applied to the text generates:
Pre-EBT -> On 7 March
Pre-EBT -> By 14 April
Pre-EBT -> in May
Pre-EBT -> In July
I -> from 24 July until 26 November

Finally the normalization of these expressions gives
On 7 March <--> 7-03-XXXX
By 14 April <--> 14-04-XXXX
in May <--> Pattern not identified yet
In July <--> Pattern not identified yet
from 24 July until 26 November <--> 24-07<X<26-11

where:
XXXX: is the unknown year.

IV. VERIFICATION AND VALIDATION

A. Verification

The verification process was responsible for answering the question about extensibility and flexibility of RISO-TT, which asks: "*Is the model of the RISO Temporal Tagger flexible and extensible?*".

To answer this question, we carried out three tests with the WikiWars corpus, with three different rules configurations, where the adjustment made in each version was based on patterns not found in the previous version.

To exemplify this process, imagine that a document *d* has a set of Temporal Expressions (TE) and that this document was marked by a Temporal Tagger (TT), resulting in a document *d'*. This document d' is composed of a set of temporal marks defined as $TM = \{m_1, m_2, .., m_n\}$, where each m_i is an expression marked with basis in the set of temporal rules $R = \{p_1, p_2, ..., p_n\}$. A temporal standard p_i is formed by a set of temporal expressions. In this case, $TM \subseteq TE$

Analyzing document *d'*, we noticed that there are temporal expressions *E'* that were not marked by *TT*; and this occurs because the rule *p* that is able to identify a temporal expressions *E'* does not belong to the set *R* of rules. That is, $p \notin R$.

Once the new rule has been inserted in the Rules file and the Configuration file has been updated, a new test has been carried out with R'=R $\cup$ {p}, obtaining a new *d''* determining *MT'* such that MT $\subseteq$ MT'.

B. Validation

In order to validate the development of RISO-TT and analyze its performance, we realized a comparative experiment. We selected two temporal marking tools for this comparison: Heideltime and SUTime. The tests were performed and the results computed and compared with the ideal markings mold, made available by WikiWars.

The WikiWars corpus has a mold with the temporal markings that exist in all documents that compose it. Based on these documents, the results of the markings by the tools chosen for experiment were compared and evaluated according to the number of correctly marked expressions, the missing expressions and those incorrectly marked.

Based on the information found, we computed the Accuracy, Coverage and F-Measure of the samples. The results are presented in Table III.

C. Data Analysis

Right after tests, the first task carried out was the evaluation of normality of the resulting data. For this activity, we ran the Shapiro-Wilk test, obtaining the results displayed in Table I.

TABLE I. SHAPIRO-WILK NORMALITY TESTS

	W	P-value
Heideltime	0.652	0,00508
SuTime	0.942	0.2176
RisoTT	0.9831	0.9574

We notice that the p-value obtained from the Heideltime data characterizes a non-normalized sample. However, evaluating the data, we noticed that they concern three samples which, in the document marking process, had not a good result. The format in question is for dates with three characters (e.g., 200 AD). This format was not recognized by Heideltime and, so, due to the non-

 ISBN: 978-1-61208-293-6

expressive number of documents, will be considered as outlier in the research.

We performed the statistical test based on the trust interval of 95%. For this, we computed the sample mean, standard deviation, and standard error. Based on this, the results found were presented in Table II.

TABLE II. TRUST INTERVALS

	Average Sample – Standard Error	Average Sample	Average Sample + Standard Error
Heideltime	0.7187674	0.7792521	0.8397369
SuTime	0.7566779	0.7842650	0.8118521
RisoTT	0.8970079	0.9139180	0.9308280

Based on the BoxPlot presented in Figure 2, it is possible to conclude that we cannot state whether there is superiority of one of the tools Heideltime and SUTime, since the trust intervals intersect each other. This occurs maybe due to the fact that both tools use the same marking standard and are based only on the knowledge based available in the standard. However, it is possible to state that there is a statistically proved superiority of the results of RISO-TT, compared to the others.

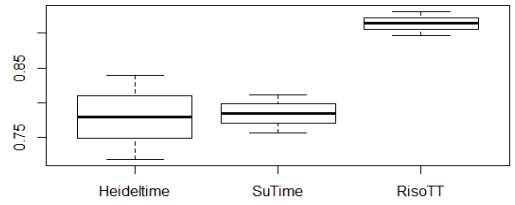

Figure 2. BoxPlot of the Trust Interval

We believe that this superiority is due to the number of relations between the mapped temporal standards and their relations, defined by the rules, in RISO-TT. The processing of information, due to this number of relations, is slower than the other tools and this may cause the making of temporal expressions to be more detailed than the others.

V. CONCLUSION AND FUTURE WORK

The development of RISO-TT is part of the RISO [18] research project and the experiments proved that it is extensible and flexible, and its performance was superior of other existing approaches.

A. Contributions

Considering the information presented throughout this paper, the main contributions of RISO-TT are:

- Flexible and extensible architecture based on standards and rules configurable by means of XML files;
- Independence of third-party software;
- Temporal Expression Recognition based on rules priority analysis;
- Possibility of creating complex structures of temporal and grammatical associations;
- Extends the standards with the possibility of arrangements and associations with other non-temporal expressions;
- Normalization of complex temporal standards, taking intervals between temporal tokens into consideration.

As future work we list:

- Concerning the temporal expression normalization process, incomplete time of an expression in a sentence may be completed by metadata about the document or other times of the current phrase or paragraph.
- The temporal expressions recognizer could be integrated with a spatial expressions recognizer.
- Recognition and treatment of ambiguities of the temporal expressions found in the document (e.g. May (Mouth) or may (Verb)).

With these documents indexing procedures integrated in the RISO project a Semantic Query Processor is under development to take into account this rich indexing structure of documents in order to optimize the quality of the information retrieval process.

B. Model Packaging

The RISO-TT Project is available in the RISO website [18].

REFERENCES

[1] R. Baeza-Yates and B. Ribiero-Neto. “Modern Information Retrieval”. Boston: Addison-Wesley Longman,1999.

[2] H. Cunningham. “JAPE: a Java Annotation Patterns Engine”. Research Memorandum CS–99–06, Department of Computer Science, University of Sheffield, May, 1999.

[3] A. X. Chang and C. D. Manning, “SUTIME: A Library for Recognizing and Normalizing Time Expressions”. 8th International Conference on Language Resources and Evaluation (LREC 2012), 2012.

[4] E. Ferneda. “Processamento da linguagem natural”. Available in: <http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/MRI-06%20-

%20Processamento%20da%20Liguagem%20Natural.pdf>. Accessed in: 12 April. 2012.

[5] K. Hacioglu,Y. Chen and B. Douglas, "Automatic time expression labeling for english and chinese text". In: GELBUKH, A. F. Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing'05, Lecture Notes in Computer Science, Mexico City, Mexico, February. Springer, 2005, pp. 548–559.

[6] H. S. Llorens. "A Semantic Approach to Temporal Information Processing (PhD Dissertation)" - University of Alicante, Departamento de Lenguajes y Sistemas Informáticos, Alicante, 2011.

[7] H. S., E. Llorens and B. Navarro. "TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2". In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010, Uppsala, Sweden, 15-16 July, 2010, pp. 284–291.

[8] I. Mani, G. Wilson, B. Sundheim, andL. Ferro. "Guidelines for Annotating Temporal Information". In Proceedings of HLT 2001, First International Conference on Human Language Technology Research, J. Allan, ed., Morgan Kaufmann, San Francisco, 2001.

[9] P. Mazur and R. Dale. "WikiWars: A New Corpus for Research on Temporal Expressions". In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, MIT, Massachusetts, USA, 9-11 October, 2010, pp. 913–922.

[10] M. Negri and L. Marseglia. "Recognition and Normalization of Time Expressions: ITC-irst ar TERN 2004 MEANING - Developing Multilingual Web-scale Language Technologies", 2004.

[11] J. Pustejovsky et al. "TimeML:Robust Specification of Event and Temporal Expressions in Text". In IWCS-5, Fifth International Workshop on Computational Semantics, Tilburg, The Netherlands, January, 2003.

[12] E. Saquete. "ID 392:TERSEO + T2T3 Transducer. A System for Recognizing and Normalizing TIMEX3". In: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010.

[13] J. Strotgen and M. Gertz, "Heideltime: High Quality Rule-based Extraction and Normalization of Temporal Expressions". In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010, Uppsala, Sweden, July 15-16, 2010, pp. 321–324.

[14] Timeml Working Group. "Guidelines for Temporal Expression Annotation for English for TempEval 2010". 2010.

[15] M. Verhagen. "Temporal closure in an annotation environment". Language Resources and Evaluation, no. 39, , May, 2005, pp. 211–241.

[16] M. Verhagen et al. "Automating temporal annotation with TARSQI". In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics, Ann Arbor, USA, 2005.

[17] C. Zambenedetti. "Extração de Informações sobre Bases de Dados Textuais". 2002. 144 f. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

[18] RISO. Avaliable in: https://sites.google.com/a/copin.ufcg.edu.br/riso-t/projeto>. [retrieved: July, 2013]

TABLE III. ACCURACY, COVERAGE AND F-MEASURE RESULTS

	Accuracy			Coverage			F-Measure		
Documents	**Heideltime**	**Sutime**	**Riso-TT**	**Heideltime**	**Sutime**	**Riso-TT**	**Heideltime**	**Sutime**	**Riso-TT**
01_WW2	0,86227545	0,8882	0,994	0,847059	0,841176	0,970588	0,854599	0,864048	0,982143
02_WW1	0,89177489	0,85	0,984	0,777358	0,769811	0,958491	0,830645	0,807921	0,971319
03_AmCivWar	0,85135135	0,7857	0,911	0,84	0,733333	0,96	0,845638	0,758621	0,935065
04_AmRevWar	0,86896552	0,8601	0,946	0,857143	0,836735	0,959184	0,863014	0,848276	0,952703
05_VietnamWar	0,84729064	0,8191	0,938	0,702041	0,665306	0,865306	0,767857	0,734234	0,900212
06_KoreanWar	0,87301587	0,7724	0,963	0,738255	0,637584	0,865772	0,8	0,698529	0,911661
07_IraqWar	0,89035088	0,8018	0,967	0,821862	0,704453	0,825911	0,854737	0,75	0,89083
08_FrenchRev	0,86363636	0,8067	0,94	0,76	0,691429	0,897143	0,808511	0,744615	0,918129
09_GrecoPersian	0,51724138	0,7477	0,893	0,232558	0,620155	0,837209	0,320856	0,677966	0,864
10_PunicWars	0,88235294	0,8085	0,922	0,263158	0,666667	0,824561	0,405405	0,730769	0,87037