Approaches for Temporal Information Extraction: A Comparative Study

Parul Patel
Assistant Professor,
M.SC(I.T) Programme,
VNSGU,Surat

Dr. S. V. Patel
Professor,
Department of Computer Science,
VNSGU,Surat

***Abstract*—In some applications like document summarization based on chronological events, question answering in newswire domain etc we need to extract and process temporal information. For example in Question Answer system in newswire domain, if we want to know who was the prime minister of India in February of 2000, we need to extract and process temporal information in the text. Extracting temporal information from text is a crucial task which needs recognition of temporal expressions, identification of events in document to which expression is related and normalization of temporal expression into a value. The information obtained may be further processed and presented to the user as per requirement. This has been an active research area especially in NLP where new tools and techniques are being developed. This paper presents a comparative study of such tools and techniques for temporal information extraction from natural language texts.**

***Index Terms*—Temporal Information Extraction, temporal Expression recognition, Temporal event**

I. INTRODUCTION

Electronic information increases rapidly day by day, so precise temporal representations of the text becomes important for many applications like text summarization, question answering , information extraction etc. The representation and interpretation of time can be done in many ways,therefore it is essential task to recognize and process it which can be used in other applications.

Temporal Information processing can be categorized in 3 steps: (1) Recognition of temporal expression i.e a task of identifying proper phrases with temporal semantics in the text. (2) Normalization of temporal expression i.e a task of assigning absolute value to temporal expression(3) Identify temporal relation between event and time expressions.

Extracting and Processing of temporal expressions has received increasing interest especially in NLP research community over past few years. The Message Understanding Conferences (MUC-6) have played a important role , but their evaluation focus only on extraction of temporal expression, while a Novel contribution towards the Normalization of temporal expressions was made in 2000**[1].** When an information is extracted, the quality of extractor is important. It is measured in terms of precision and recall as per standards. The first exercise evaluating system performance that deals both with recognition and

normalization of temporal expression was the TERN 2004 competition. There after intensity of research activities in this promising area has increased substantially.

This paper provides comparative study of various approaches, tools and techniques of temporal information extraction.

II. TEMPORAL EXPRESSIONS IN DOCUMENTS

Temporal expressions can be classified into following categories according to Schilder and Habel[2].

Explicit: Date Expressions such as '13/08/2013','15th August' refer explicitly to entries of a calendar system and can be mapped directly to Chronons in a timeline.

Implicit: All temporal expressions that can be evaluated via a given time ontology and capability of the named entity extraction approach such as name of holiday(last chritmas), next valentine day etc.

Relative : Some temporal expressions express vague temporal information and it is rather difficult to precisely place the information expressed on a time line. Such temporal expressions can be only anchored in a timeline in reference to another explicit or implicit already anchored temporal expressions. For example, ' on Monday','Before June and After March' etc. If the document has creation date, then they can be easily anchored.This date then can be used as a reference for that expression which can be then mapped to a chronon.

II. Annotation standards

There are two standards for annotating temporal information in documents: TIMEX2 [3]and TIMEML[4].

TIMEX2

TIMEX2 was initially developed in 2000 under DARPA's Translingual Information Detection, Extraction and Summarization (TIDES) program. The fundamental move forward here (compared to TIMEX) was the addition of a normalisation task to the recognition task: the annotation provided for the interpretation of dates and times by using ISO

8601 as the standard for the representation of normalized dates and times.Various attributes with their meanings are as shown in table I.

TABLE I
TIMEX2 ATTRIBUTES

Attribute	Description
VAL	Contains a normalized form of the date or time in the annotated expression.
MOD	Captures temporal modifiers, using values such as BEFORE, MORE_THAN, START, and APPROX.
ANCHOR_VAL	Contains a normalized form of an anchoring date or time.
ANCHOR_DIR	Captures the relative direction or orientation between VAL and ANCHOR_VAL attributes, as in WITHIN, STARTING and BEFORE. It is used to express the information about when a duration is placed.
SET	Identifies expressions denoting sets of times, either takes the value YES or is empty.
COMMENT	Contains any comment that the annotator wants to add to the annotation, ignored from the point of view of automatic processing of the text.

TIMEML

TimeML is a robust specification language for events and temporal expressions in natural language [5]. There are four major data structures that are specified in TimeML: EVENT, TIMEX3, SIGNAL, and LINK

It has been applied mainly to English news articles. It is designed to address four problems in event and temporal expression markup:

1. Time stamping of events (identifying an event and anchoring it in time);
2. Ordering events with respect to one another (lexical versus discourse properties of ordering);
3. Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before');
4. Reasoning about the persistence of events (how long does an event or the outcome of an event last).

TimeML [5] is a proposed metadata standard for markup of events and their temporal anchoring in documents that addresses this.. Both standards present guidelines on how to determine the extents and how to normalize the values of temporal expressions. TimeML contains following attributes as shown in table II.

TABLE II
TIMEML ATTRIBUTES

Attribute	Description
<EVENT>	The EVENT tag is used to annotate those elements in a text that mark the semantic events described by it. Syntactically, EVENTs are typically verbs, although event nominals, such as "crash" in "...killed by the crash", will also be annotated as EVENTs.
<MAKEINSTANCE>	MAKEINSTANCE is a realization link; it indicates different instances of a given event. Since different instances can have different attribute values, the tense and aspect of the event are represented within this tag. In addition, if the instance is modified by a negation or modal operator, this is represented in the appropriate attributes within this tag.
<TIMEX3>	The TIMEX3 tag is primarily used to mark up explicit temporal expressions, such as times, dates, durations, etc.
<SIGNAL>	SIGNAL is used to annotate temporal functions words like "after", "during", etc.
<TLINK>	TLINK is a temporal link; it represents the relation between two temporal elements.
<SLINK>	This is a subordination link that is used for contexts involving modality, evidentials, and factives. An SLINK is used in cases where an event instance subordinates another event instance type. These are cases where a verb takes a complement and subordinates the event instance referred to in this complement.
<ALINK>	ALINK is an aspectual link; it indicates an aspectual connection between two events. In some ways, it is like a cross between TLINK and SLINK in that it indicates both a relation between two temporal elements, as well as aspectual subordination

III. ANNOTATED CORPORA AND MEASURES FOR EVALUATION

Gold standards for evaluation of temporally annotated resources are very limited in general domain [6]. Further, in specific domains like medical, clinical and biological[7], such standards are quite less. In the last decade, different sources of annotated temporal expressions have been developed, but all of them have used different annotation standard.The major difference is that they have used different tags to annotate temporal expressions. Some of them are tagged according to TIMEX2 and some are as per TIMEX3. Due to these differences, it becomes impossible to use both the corpora at the same time for the same annotated temporal expression .The TimeBank 1.2 Corpus contains 183 news articles that have been annotated following the TimeML 1.2.1 specification [8]. WikiWar is corpus of temporally rich documents sourced from English Wikipedia which have been annotated with TIMEX2 tags [9]. It contains around 12000 tokens, and 2600 TIMEX2 expressions. Michele [10] has created a corpus of temporal expressions collecting all TIMEX3 tags in four different corpora: AQUAINT, TimeBank 1.2, Wikiwars and TRIOS Time Bank v0.1.

IV.QUALITY AND PERFORMANCE MEASURES

PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

$$precision = \frac{|\{relevant\ expression\} \cap \{retrieved\ expressions\}|}{|\{retrieved\ expressions\}|}$$

RECALL is the ratio of the number of relevant records retrieved to the total number of relevant records in the database which is expressed as a percentage.

$$recall = \frac{|\{relevant\ expression\} \cap \{retrieved\ expressions\}|}{|\{relevant\ expressions\}|}$$

F-measure

F measure is mean of precision and recall. The balanced F-score or traditional F-measure is:

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

As recall and precision are weighted evenly in this measure, it can also be considered as F1 measure.

V. APPROACHES FOR TEMPORAL PROCESSING

There are three basic approaches for temporal information processing.

(I) Rule Based Approach:

Initially rule based system for information extraction was implemented. Such systems are based on domain specific extraction rules written by a domain skilled person. Several rule based systems were developed to extract information namely chromos [11],AutoSlog by Riloff[12]; Fastus [13] by Appelt et, al. and GATE by cunnungham et al[14].These methods require a great human effort and a considerable time for data analysis and rule writing. Also rule based methods lack portability to other domains.

Chronos is a rule based system designed with the aim to carry out recognition as well as normalization of temporal expressions. It annotate textual data by using TIMEX2 tag which contains attributes for expressing the normalized value. Basic and Composition rules are defined for recognition and normalization phase.

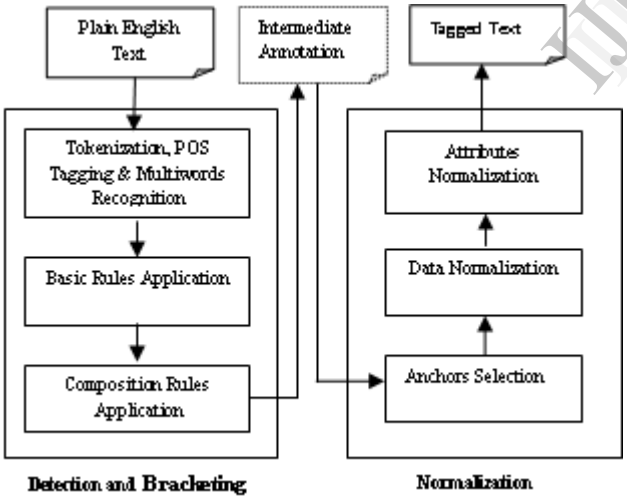

Figure 1: The architecture of Chronos [11].

SUTime[15] has applied rule based approach for recognition of temporal expression. They have defined three different set of rules. (a) text regular expression rules, which map regular expression over a character or token to temporal representations (b) compositional rules for chunk of tokens (c) Filtering rules for removing ambiguity in temporal expressions. The DANTE[16]temporal tagger has recognized temporal expressions using JAPE grammar. It contains five phases which runs over a document in sequence. Each phase contains rules which match annotations introduced by earlier processing components. To recognize and normalize temporal expression in Serbian text, Jelena[17] used finite state transducer methodology. Local grammars are designed to recognize calendar dates, time of day, periods of time and durations to determine the extensions of detected expressions as well as to normalize their values in ISO format . Hiedeltime ,a rule based system developed for extraction and normalization of temporal expressions. It uses linguistic clues for their normalization [18].

As the rule based approaches required substantial human effors, there was a need to develop learning components so that human efforts could be reduced. This inspired the development of machine learning approaches.

(II) Machine Learning System

These approaches have become the good technique for many solving problems in Natural Language Processing.For the purposes like named entity recognition, parsing, tagging and semantic role labeling, several models were used. Some of them are: (i) Generative models based on Hidden Markov Models[19,20,21] (ii)Conditional Random Field Models based maximum entropy [22,23,24] (iii) Global conditional models which use Conditional Random Fields[25].Machine learning approaches can used by (i) Dividing into token by token classification followed by B(egin), I(nside),O(utside) encoding (ii) Binary constituent based classification, in which an entire chunk phrase is under consideration to be classified as a temporal expression or not [26]. In KUL[27], recognition of temporal expression is done by using Hidden Markov Model and Latent word language model. Normalization of temporal expression is done by using rule based approach. For value estimation, they have used fixed values for vocabulary and/or executing instructions or methods specified. A set of rules is provided and can be extended with new implementations of resolution strategies. For resolution of complex relational expressions, they have applied fixed strategy i.e all relative expressions were resolved by using document creation time.Using document creation time as reference date sometimes generates error with wrong output. In a system developed by David [28] they have used support vector machine for classification. Training has been given on TERN corpus.

(III) Hybrid approach:

This third category of approaches leverages the advantages of machine learning and rule based methods[29,30,31]. Several researchers have expanded the capabilities of existing Natural Language Processing using hybrid approach. TRIPS and TRIOS are hybrid system which uses deep semantic parsing, Markov logic network and Conditional Random Field Classifier. They have used tokenwise classification for temporal expressions.Further it is represented by B(Begin)-I(Inside)-O(Outside) encoding by using set of lexical and syntactic features using CRF classifier.The temporal expressions that are suggested by CRF based system, are passed to a filtering step that tries to extract a normalized value and type of temporal expression[31]. Mani[32] has developed an algorithm for temporal annotation specifications which resolves a class of temporal expressions found in news

article. This algorithm uses a mix of hand crafted rules and machine learnt rules and obtained reasonable results.

VI. COMPARASION OF VARIOUS TOOLS & TECHNIQUES

Various tools have been developed based on above approaches. We have studied and analiyzed these tools and tabulated their various aspects covering standard, methods for recognition and normalization, quality and performance measures such as precision, recall and F-score as shown in Table III:

TABLE III

Tool	Standard	Recognition	Normalization	Precision	Recall	F-score	TYPE	VAL
KUL		Supervised Machine Learning trained on TimeBank	Rule based technique	0.85	0.84	0.845	0.91	0.88
DANTE	TIMEX2	Rule Based technique	Rule based technique	59.0	57.2	58.6	-	-
UCSM	TIMEX3	Rule based technique	Rule based technique	0.90	0.87	0.88	0.91	0.88
TE for Serbian Texts	TIMEX3	Finite State Automata	Finite State Automata	0.93	0.96	0.94	-	-
Chronos	TIMEX2	Rule Based Technique	Rule based Technique	0.97	0.88	0.92	0.87	0.87
TRIPS	TimeML	deep semantic parsing, Markov Logic Networks and Conditional Field Classifiers.	combination of deep semantic parsing, Markov Logic Networks and Conditional Field Classifiers.	0.55	0.88	0.67	-	-
TRIOS	TimeML	combination of deep semantic parsing, Markov Logic Networks and Conditional Field Classifiers.	combination of deep semantic parsing, Markov Logic Networks and Conditional Field Classifiers.	0.80	0.74	0.77	-	-
GUTime	Timex3	Rule Based Technique	Rule Based Technique	0.89	0.79	0.84	0.91	0.88
HeidelTime 1	TimeML	Rule Based Technique	Rule Based Technique	0.90	0.82	0.86	0.96	0.85
HeidelTime 2	TimeML	Rule Based Technique	Rule Based Technique	0.82	0.91	0.86	0.92	0.77
SUTime	TIMEX3	Rule Based Technique	Rule Based Technique	0.89	0.79	0.84	0.96	0.82
System by Olsen and Kolomiyets	TIMEX3	Machine Learning	Machine Learning	0.872	0.836	0.852	-	-

It can be seen from the above table that some systems have used machine learning approach. This system have low precision and F-Score. Rule based technique has given good result as compared to machine learning approach. Rule based technique is good enough to extract regular temporal expression, but it needs great human efforts in data analysis and rule writing. However by using a rule based technique, it is difficult to recognize ambigious expressions because it is dependent on domain expertise. Therefore we advocate to extract basic temporal expression by using rule based technique and then after find ambigious expression using machine learning methods. This would increase overall performance of the system. The proposed architecture is shown in Figure II.

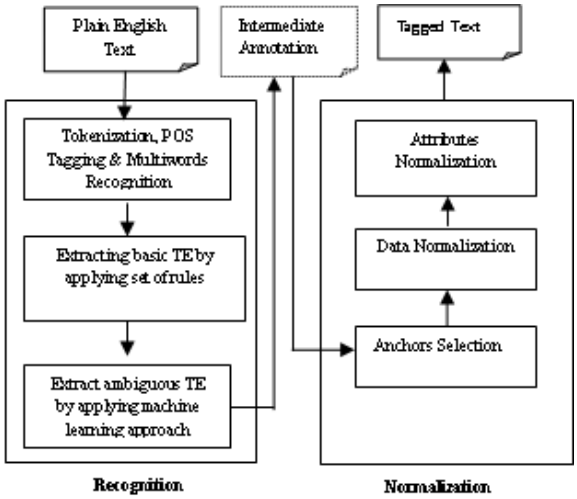

Figure II. Proposed Architecture for Temporal information Processing

VI.CONCLUSION

It can be concluded that quality and performance of Temporal information extraction can be improved by using both the techniques in proper combination and order. Further, it is also observed that most of the existing systems are designed and implemented for particular domains. (for example Newswire). Hence, there is need of efforts to be done for developing generalized model for all type of documents rather than relying on specific domain.

REFERENCES:

1. Mani I. and Wilson, G. Robust Temporal Processing of News. In proceedings of 38th Annual Meeting on Association for Computational Linguistics(Hong Kong 2000), 69-76.
2. Frank Schilder and Christopher Habel: 'From Temporal Expression to Temporal Information:Semantic tagging of News Messages' : In Proceeding of the ACL2001 Workshopon Temporal and Spatial Information Processing, 2001.
3. Laurie Gerber, Lisa Ferro, Inderjeet Mani, Beth Sundheim, and George Wilson: 'TIDES 2005 Standard for the Annotation of Temporal Expressions' Technical Report, The MITRE Corporation,2005.
4. James Pustejovsky, Jessica Littman, Robert Knippen, and Roser Sauri:'Temporal and Event Information in Natural Language Text, Language Resources and Evaluation',39(2-3):123-164,2005
5. TimeML Annotation Guidelines: Version 1.2.1 Roser Saur´ı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky: January 31, 2006
6. L.Derczynski, H.Llorens, and E. saquete. Massively Increasing TIMEX3 Resources : A Transduction Approach. ArXiv e- prints, Mar.2012.
7. L. Galescu and N.Blaylock. A corpus of clinical narratives annotated with temporal information. In proceeding of the 2nd ACM SIGHIT International health Informatics Symposium, IHI'12 ages 7150720, New York, NY,USA,2012.ACM.
8. www.timeml.org
9. WikiWars: A new corpus for Research On Temporal Expression: Pawel Mazur, Robert Dale: Proceeding of the 2010 conference on Empirical Methods in Natural Language Processing, pages, 913-922, MIT ,Massachusetts, USA,9-11 October 2010.
10. Temporal Expression Normalization in natural language texts: Michele :School of Computer Science, The university of Manchester , UK.
11. Matteo Negri and Luca Marseglia:'Recognition and Normalisation of Time Expressions: ITC-irst at TERN 2004",February 2005.
12. E. Riloff, (1993)" Automatically constructing a dictionary for information extraction tasks". In Proceedings of the 11th National Conference on Artificial Intelligence, pp 811–816.

13. D. E. Appelt, J. R. Hobbs, J. Bear, D.J. Israel & M. Tyson, (1993). "Fastus: A finite-state processor for information extraction from real-world text," in IJCAI, pp 1172–1178.
14. Cunningham, H.D. Maynard, K. Bontcheva, and V.Tablan, 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the ACL.
15. SUTime: A Library for Recognizing and Normalizing Time Expressions: Angel X.Chang, Christhpher D.Manning
16. The DANTE Temporal Expression Tagger: Pawel Mazur, Robert Dale:Wroclaw University of Technology.
17. Recognition and normalization of temporal expressions in Serbian texts: Jelena Jacimovic:BCI'12, September 16-20,2012, Novi sad, Serbia.
18. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions: Jannik Str¨otgen, Michael Gertz: Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 321– 324,Uppsala, Sweden, 15-16 July 2010
19. E. Agichtein & V. Ganti, (2004). "Mining reference tables for automatic text segmentation". In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA.
20. K. Seymore, A. McCallum & R. Rosenfeld, (1999) "Learning Hidden Markov Model structure for information extraction," in Papers from the AAAI- 99 Workshop on Machine Learning for Information Extraction, pp 37–42.
21. Zhang, N. R, (2001) "Hidden Markov Models for Information Extraction".
22. A. Borthwick, J. Sterling, E. Agichtein & R. Grishman, (1998)"Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in Sixth Workshop on Very Large Corpora New Brunswick, New Jersey, Association for Computational Linguistics.
23. A. McCallum, D. Freitag & F. Pereira, (2000) "Maximum entropy markov models for information extraction and segmentation," in Proceedings of the International Conference on Machine Learning (ICML- 2000), pp 591–598, Palo Alto, CA.
24. R. Malouf, (2002) "A comparison of algorithms for maximum entropy parameter estimation," in Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002), pp 49–55.
25. F. Peng & A. McCallum, (2004) "Accurate information extraction from research papers using conditional random fields," in HLT - NAACL, pp 329–336.
26. Meeting TempEval-2: Shallow Approach for Temporal Tagger: Oleksandr Kolomiyets, Marie-Francine Moens: Proceeding of the NAACL HLT Workshop on semantic Evaluations: Recent Achievements and Future Directions , pages 52-57. Boulder,Colorado, June 2009.
27. KUL: Recognition and Normalization of Temporal Expressions:Oleksandr Kolomiyets,Marie Francine Moens: In the Proceedings of the 5th Workshop on semantic Evaluation, ACL 2010 pages 325-328.
28. David Ahn, Joris van Rantwijk, Maarteen de Rijke :A cascaded Machine Learning Approach to interpreting temporal expressions
29. M. Califf & R. Mooney, (2003) "Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction".
30. B. Marthi, B. Milch, & S. Russell, (2003) "First-order probabilistic models for information extraction," in Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL- 2003), (L. Getoor and D.Jensen, eds.), pp 71–78, Acapulco, Mexico.
31. Y. Choi, C. Cardie, E. Riloff, & S. Patwardhan, (2005) "Identifying sources of opinions with conditional random fields and extraction patterns," in HLT/EMNLP.
32. [31] Naushad UzZaman :'TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text':
33. University of Rochester, Rochester, NY, USA.
34. [32] Frank Schilder and Christopher Habel: 'From Temporal Expression to Temporal Information:Semantic tagging of News Messages' : In Proceeding of the ACL2001 Workshop on Temporal and Spatial Information Processing, 2001.