Under consideration for publication in Knowledge and Information Systems

Normalization of Imprecise Temporal Expressions Extracted from Text

Hegler Tissot[1], Marcos D. Del Fabro[1], Leon Derczynski[2] and Angus Roberts[2]

[1]Department of Computer Science, Federal University of Paraná, Curitiba PR, Brazil
[2]Department of Computer Science, The University of Sheffield, Sheffield, UK

Abstract.

Information extraction systems and techniques have been largely used to deal with the increasing amount of unstructured data available nowadays. Time is amongst the different kinds of information that may be extracted from such unstructured data sources, including text documents. However, the inability to correctly identify and extract temporal information from text makes it difficult to understand how the extracted events are organised in a chronological order. Furthermore, in many situations, the meaning of temporal expressions (timexes) is imprecise, such as in "less than 2 years" and "several weeks", and cannot be accurately normalised, leading to interpretation errors. Although there are some approaches that enable representing imprecise timexes, they are not designed to be applied to specific scenarios and are difficult to generalise. This paper presents a novel methodology to analyse and normalise imprecise temporal expressions by representing temporal imprecision in the form of membership functions, based on human interpretation of time in two different languages (Portuguese and English). Each resulting model is a generalisation of probability distributions in the form of trapezoidal and hexagonal fuzzy membership functions. We use an adapted F1-score to guide the choice of the best models for each kind of imprecise timex, and a weighted F1-score ($F1_{3D}$) as a complementary metric in order to identify relevant differences when comparing two normalisation models. We apply the proposed methodology for three distinct classes of imprecise timexes and the resulting models give distinct insights in the way each kind of temporal expression is interpreted.

Keywords: Natural Language Processing (NLP); Information Extraction; Temporal Expression (Timex); Imprecise Timex Normalisation

Received xxx
Revised xxx
Accepted xxx

1. Introduction

The extraction of temporal information from text is fundamental for language understanding (Burman, Jayapal, Kannan, Kavilikatta, Alhelbawy, Derczynski and Gaizauskas, 2011) and an important sub-task for several language processing applications (UzZaman and Allen, 2010), such as text summarisation and knowledge base population. Processing a temporal expression (timex) from text, i.e. extracting and modelling the expression, includes tasks such as recognition and representation of the temporal information (Kolomiyets, 2012). Solving challenging computational problems involving time has been a critical component in the development of information extraction (IE) systems (Bartak, Morris and Venable, 2013), e.g., understanding how such elements that describe temporal concepts can be formally represented and what procedures should be performed by an algorithm to deal with the set of operations that we as humans seem to perform relatively easy (Caselli, 2009).

In many situations, however, extracted temporal expressions are not accurately described in the text, i.e. the expressions denote an imprecise amount or point in time, as in "about 3 months ago", "less than a year", "few days", and "recently". More than 30% of temporal information in some text types, e.g. clinical notes, can be imprecise, affecting for example the results of searches for events related to such temporal data. In addition, an inaccurate interpretation may yield different values for the same expression. For this reason, for a given application, it is important to estimate standardised values for the existing imprecise timexes, i.e., to normalise them.

TimeML (Pustejovsky, Castano, Ingria, Saurí, Gaizauskas, Setzer and Katz, 2003) is the major initiative for temporal information annotation being an ISO standard since 2010. It is designed to connect the processes of temporal analysis of a text with a representation and formal meaning of time, providing a model and annotation scheme for temporal information in text, including the TIMEX3 scheme for representing temporal expressions. Although TimeML is capable of describing imprecise timexes in terms of language structure, it does not provide mechanisms to correctly normalising them. Therefore, the normalisation of imprecise temporal data in terms of values can be ambiguous or incomplete, e.g. it provides one `mod` attribute that allows the modification of expressions, but only in a very constrained way (twelve preset non-disjoint modifiers). In order to overcome this lack, existing approaches (Nagypál and Motik, 2003; Schockaert, 2005; Schockaert, Cock and Kerre, 2008; Filannino and Nenadic, 2014) use fuzzy sets to represent individual timexes and relations. However, they describe specific historical events or generic periods of time (e.g. holidays), relying on external sources of data, such as the result of Internet search queries or image timestamps collected from social media, and they do not provide a generic or reusable methodology for the normalisation of imprecise timexes. In these situations, the normalisation is done based on the extracted time spans, which are often focused in one kind of expression and with restricted interpretation of the timexes, being difficult to be applied to broader domains.

This paper contributes with an analysis of a previously unstudied set of imprecise temporal expressions, and presents a novel method for their normalisation and representation. The main contributions are the following:

Imprecise timexes quantification and classification. The classification was done based on the expressions extracted from clinical narratives. This classification is used as basis for the presented approach.

Methodology for imprecise timex normalisation. We introduce a novel methodology for the normalisation of imprecise temporal expressions extracted from text. Our methodology comprises a set of steps, starting from creating a set of questionnaires used to capture how people interpret vague descriptions of time in text. The questionnaires were designed from scratch, since there is not a dataset or standard for evaluation of imprecise timexes. Answers were used as input data, from which we created histograms and fuzzy membership functions (MSF) during the pre-processing step. Then, we applied statistical regression and machine learning (ML) techniques in order to evaluate which would be the most suitable model for each kind of temporal imprecision being evaluated. The result is a grounded probability density function for the period over which the timex was attained. We use F1-score to calculate how similar two membership functions are, and to choose the most suitable representation model for each kind of imprecise temporal expression.

Weighted F1-score. We presented a new weighted F1-score variation, called $F1_{3D}$, that better identifies the relevant differences between two membership function in terms of confidence, by checking whether the differences are more concentrated in the top or in the bottom when comparing two membership function shapes or two normalisation models. We apply the presented methodology for three kinds of imprecise timexes, and we compare the normalisation models results in English and Portuguese. The results showed that the normalisation models were able to capture the vagueness carried out by the imprecise timexes.

This paper is organised as follows: Section 2 presents the background and related work regarding to the temporal information extraction and the normalisation of imprecise timexes; Section 3 presents a quantification of imprecise expressions comparing clinical and non-clinical domains, and proposes a classification for imprecise timexes; In Section 4, we propose a methodology for the normalisation of imprecise temporal expressions; Section 5 depicts the normalisation models resulted for three types of imprecise expressions and compares the normalisation models for two different languages (Portuguese and English); lastly, in Section 6, we present the final conclusions and future work.

2. Background and Related Work

Time is a primary element that allows us to observe, describe and reason about what surrounds us in the world, providing a substrate for the human management of perception and action. As a cognitive and linguistic component for describing changes which happen through the occurrence of events, processes, and actions, time provides a way to record, order, and measure the duration of such occurrences (Bartak et al., 2013). As a pervasive element of human life, the absence of a correct identification of the temporal ordering may result in a bad comprehension, leading to a misunderstanding (Caselli, 2009).

2.1. Temporal Information Extraction

The general process of reading and understanding a text includes the inference about whether the presented situations stand in particular points in time (Caselli, 2009). Organising events in a chronological order is important to find the temporal relations (e.g. before/after relations) amongst them. Temporal information

extraction plays an important role in this respect. Temporal expressions are written in natural language and can refer directly to time points or intervals (e.g. "6 years ago"), serving as anchors for linking concepts and events extracted from the text to a timeline, providing the correct distribution of such extracted elements in time (Ahn, Adafre and Rijke, 2005). Nevertheless, this seemingly easy task takes into account a set of complex information involving different linguistic entities and sources of knowledge (Caselli, 2009).

The recognition (or annotation) of temporal expressions (timexes) in text is the task of finding the corresponding labels $(y_1, ..., y_n)$ to a given input string of tokens $(x_1, ..., x_n)$ so that the resulting labelling can be decoded into textual spans that constitute the tokens and denote time in the input string (Kolomiyets, 2012). According to Fagerberg (2014), the temporal information extraction process comprises: a) temporal expressions have to be recognised within some kind of document and extracted from it; and b) extracted temporal expressions should be categorised and normalised to a canonical form – normalisation is not just a formatting problem, but a task in which the appropriate value of the extracted expression has to be calculated.

TimeML[1] (Pustejovsky, Castano, Ingria, Saurí, Gaizauskas, Setzer and Katz, 2003) became a ISO[2] standard in 2010, as a language for temporal information annotation, designed to connect the processes of temporal analysis of a text with a representation and formal meaning of time. As a specification language for event and temporal expressions in natural language text, TimeML is able to capture distinct phenomena in temporal markup.

Temporal information extraction approaches are usually focused on recognising temporal expressions in text, and normalising those expressions by using a function that transforms the matched expression into a normalised form based on <TIMEX3> tags (Bethard, Martin and Klingenstein, 2007; Fagerberg, 2014). In Llorens, Derczynski, Gaizauskas and Saquete (2012), authors use the argument that temporal expression normalisation can only be effectively performed with a large knowledge base and set of rules.

The TempEval series in SemEval (International Workshop on Semantic Evaluation) have been exploring the task of extracting temporal expressions, events, and temporal relations from text, with the purpose to advance research on temporal information processing. SemEval-2015 Task 6 Clinical TempEval[3] (Bethard, Derczynski, Pustejovsky and Verhagen, 2015) and SemEval-2016 Task 12 Clinical TempEval (Bethard, Savova, Chen, Derczynski, Pustejovsky and Verhagen, 2016) were temporal information extraction tasks over the clinical domain, using clinical notes and pathology reports for cancer patients. Results of TempEval-3[4] and Clinical TempEval (2015[5] and 2016[6]) were given in terms of Precision, Recall and F1-score (Davis and Goadrich, 2006) relevance measures.

In addition to SemEval TempEval series, the i2b2 Natural Language Processing Challenge for Clinical Records (Sun, Rumshisky and Uzuner, 2013) focused on the temporal relations in clinical narratives, attracting 18 participating teams

[1] http://timeml.org/
[2] https://www.iso.org/standard/37331.html
[3] http://alt.qcri.org/semeval2015/task6/
[4] https://www.cs.york.ac.uk/semeval-2013/task1
[5] http://alt.qcri.org/semeval2015/task6/index.php?id=results
[6] http://alt.qcri.org/semeval2016/task12/index.php?id=results

to analyse discharge summaries, annotating time expressions, events, and relations between them.

2.2. Normalisation of Temporal Expressions

Normalisation of temporal expressions (or Timex Normalisation) is the process of tagging a timex, by setting attribute values that describe that expression in terms of an amount of time or a point in time (Kolomiyets and Moens, 2010). The timex normalisation task consists of obtaining the absolute value of a timex regardless of the linguistic expression used (Llorens et al., 2012). After a timex is recognised, its temporal value must be defined, which means finding the value attribute for such temporal expression. The normalisation process is usually implemented as a rule-based system to overcome some problems, including: a) the infinite number of possible labels, and b) the large number of ways a calendar value can be expressed in natural language (Kolomiyets, 2012).

Current annotation standards are restricted to normalise imprecise timex in terms of language structure or language elements (Ferro, Gerber, Mani, Sundheim and Wilson, 2005; Sauri, Littman, Gaizauskas, Setzer and Pustejovsky, 2006; Pustejovsky, Lee, Bunt and Romary, 2010; Styler, Bethard, Finan, Palmer, Pradhan, de Groen, Erickson, Miller, Lin, Savova and Pustejovsky, 2014). An expression like "few weeks" is normalised to represent an "undetermined period of time" or an "undetermined number of weeks", making it hard to connect that expression to a timeline without any numerical value. When improving the normalisation guidelines to consider a timex description in terms of uncertain values or periods of time (e.g. range of values), events related to imprecise timexes can be chronologically placed, and temporal reasoning can be applied.

Although it is relatively easy to recognise temporal expressions using rule-based systems or supervised machine learning approaches, normalisation (interpreting them accurately) is a complex task that requires human knowledge, since any practical approach to timex normalisation requires a hand-crafted rule set (Llorens et al., 2012). Kolomiyets (2012) presents a TimeML-based normalisation technique that comprises three sub-tasks:

1. Timex classification: a classifier has to distinguish between 4 different labels of DATE, TIME, DURATION and SET, to define the type of time expression, as it is defined in TimeML; a rule-based method performs the semantic analysis of time expression constituents (token labelling), identifying different categories (Table 1) with a comprehensive vocabulary and a set of context dependent normalisation rules specific for that category.
2. Estimation of temporal values: temporal values are estimated (normalised); this is not considered a difficult task for absolute temporal expressions, because such kinds of timexes contain all components required for calculating the final value. Relative expressions ("last week", "next month") also can be represented using ISO standards (ISO, 2007) representation facilities.
3. Aggregation of temporal values: an aggregation of temporal values is performed, when one temporal expression consists of a set of shorter temporal expressions that are obtained by pre-normalisation; in this case, partially estimated values are aggregated to obtain a final temporal value.

Table 1. Timex categories (Kolomiyets, 2012).

Category	Examples
Temporal units	day, month, year
Temporal modifiers	last, previous, next
Temporal quantifiers	several, few
Temporal directions	ago, further, later
Temporal approximators	almost, about
Day names	Monday, Tuesday
Month names	January, February
Cardinal numbers	one, 1, two, 2
Ordinal numbers	first, 1st, second, 2nd
Coreference timex	period, time
Fixed timex	today, yesterday, now

2.3. Imprecise Temporal Representation

Considerable effort has been carried out to extract temporal information from natural language texts, allowing question answering systems to deal with more complex temporal questions. However, temporal relationships expressed in natural language are often vague (which is inherently associated with real-world temporal information), and it is necessary to extend traditional temporal reasoning formalisms to cope with this kind of vagueness (Schockaert et al., 2008).

In temporal question answering systems, answering a complex question may require decomposing the original question into partial questions, to answer such partial questions and combine the partial answers into the final answer. Temporal questions are an important class of complex questions, in which the accurate representation of the time span of events is essential to the treatment of such complex questions (Schockaert, 2005).

However, a lot of time information is ill-defined, subjective or uncertain, and the boundaries of time periods can often be vague. Thus, the time span representation should be tolerant of imprecision in temporal question answering systems. Zhou, Li, Lu and Duan (2011) summarised the common types of temporal expressions, based on an exhaustive analysis of 147 clinical records, establishing temporal expression classification from such expressions. Despite including uncertain temporal expressions in the resulted classification, the authors state that the automatic extraction work was hampered by the existence of such expression type.

Although TimeML is able to distinguish imprecise temporal expressions, it is restricted to describe imprecision in terms of language structure, clouding later temporal processing. For example, in the sentence "frequent headaches for less than one month", a patient tries to describe how long a headache has lasted. The corresponding amount of time, however, cannot be accurately defined, due to the modifier "less than". The target imprecise expression "less than one month" is annotated in TimeML as <TIMEX3 value="P1M" mod="LESS_THAN">. As a consequence, when interpreting this expression and its annotated features, it is not clear whether we should consider each possible number of days between 0 and 30 as equally likely, or whether for example, 20-25 days ago is more likely than 5-10 days ago or even "yesterday".

The fuzzy set theory is a representation formalism suitable for this purpose, allowing the definition of a gradual beginning and ending of events (Nagypál and

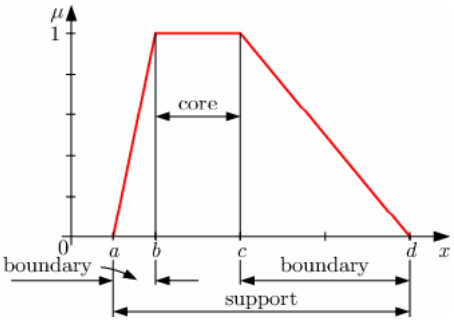


Figure 1. Concepts related to a fuzzy set (Coelho and Raposo, 2005).

Motik, 2003). A fuzzy set is the basic concept that underlies the fuzzy systems theory (Pedrycz and Gomide, 1998), and involves capturing, representing, and working with linguistic notions, being employed in those circumstances where impreciseness, unpredictability, and vagueness are in concern. A fuzzy set S is characterised by a membership function A mapping the elements of a (finite or not) domain, space or universe of discourse T into the unit interval $[0, 1]$. That is, $A(t) : T \rightarrow [0, 1]$ (Zadeh, 1994). A membership function A can be defined in different forms, such as triangular or trapezoidal functions, or continuously differentiable curves with smooth transitions, such as normalised Gaussian functions. The *height* of a fuzzy set S is the largest membership grade of any element in that set (Equation 1), whereas a fuzzy set S is called *normal* when $height(S) = 1$, and *subnormal* otherwise (Pedrycz and Gomide, 1998).

$$height(S) = max\,\{A(t), t \in T\} \tag{1}$$

The *support* of S, $supp(S)$, is the crisp set with all the elements of T satisfying $A(t) > 0$. Likewise, the *core* of S, $core(S)$, is the crisp set with all the elements of T satisfying $A(t) = 1$, whereas its *boundary*, $bound(S)$, encompasses all the elements of T with membership grades in the range $]0, 1[$, as shown in Figure 1 (Coelho and Raposo, 2005).

Although some proposed approaches and systems can identify temporal information in text (Kolomiyets and Moens, 2013; Chambers, 2013; Bethard, 2013; Strötgen, Zell and Gertz, 2013), they do not deal with imprecise temporal expressions, like "a few weeks ago" or "the coming months", in terms of defining more specific attributes to describe and connect those expressions to a timeline. Such approaches do not implement temporal-related logics to manipulate such inaccurate information, for example, to compare events associated respectively to expressions such as "about 2 months ago" and "a few weeks ago", indicating which one happened before or after (Ling and Weld, 2010).

In Nagypál and Motik (2003), a fuzzy interval-based temporal model capable of representing imprecise temporal knowledge is described. It generalises Allen's (Allen, 1983) temporal relations on intervals, by providing a definition of crisp interval relations based on set theory and then generalised them to the fuzzy case. The presented temporal model is intended for use in ontology modelling, following a modular semantics pattern which tries to keep the semantics of each model separate and to provide clean interfaces between them. Examining

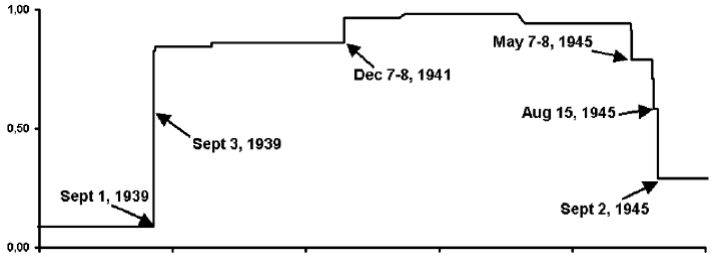


Figure 2. Fuzzy set representing the time span of World War 2 (Schockaert, 2005).

the different properties of the fuzzy temporal relations (like transitivity), one can observe basic inferences even in case of fuzzy intervals.

Schockaert et al. (2008) presents a framework to represent, compute and reason about temporal relationships between events that have imprecise time spans, represented by fuzzy sets (*fuzzy time intervals*). The proposed model preserves many of the Allen's relations' properties, and it uses a transitivity table for efficient fuzzy temporal reasoning. The qualitative relations between two fuzzy intervals are defined in terms of the ordering of the gradual beginning and endings of these intervals (ordering of the time points belonging to these intervals). It also defines four basic fuzzy relations to order two time points a and b (long before, before or at approximately at the same time, approximately at the same time, just before). Four basic fuzzy relations are defined to order two time points a and b (long before, before or at approximately at the same time, approximately at the same time, just before).

Schockaert (2005) suggests an approach based on fuzzy sets to define the beginning and ending of events, and provides a fully automatic procedure which uses statements on the web to construct the membership functions. To obtain useful statements from the web, authors used the snippets returned by Google[7] for some automatically generated queries. In most applications, all membership functions are defined by an expert. However, this is considered the first attempt to construct membership functions for fuzzy time periods in an automatic way. Figure 2 shows an example that considers the time span of the World War 2. There does not exist a unique point in time that corresponds to the beginning or ending of this war.

A similar approach was used in Blamey, Crick and Oatley (2013) to represent a temporal expression S by a function $f(t)$, which is a probability density function for the continuous random variable T_s, using photographs uploaded to the photo-sharing site Flickr.[8] After collecting a list of timestamps for an specific temporal term, the target is to find a probability density function to provide a convenient representation, and smooth the data appropriately. Authors argue that temporal expressions can communicate more than points and intervals, and their cultural meaning is much more complex – often difficult to be precisely defined. Thus, a distributed definition can capture such cultural meaning in a more detailed way, as shown in Figure 3 for the expression "Christmas".

Even though the related work described uses fuzzy sets to represent individual

[7] `http://www.google.com`
[8] `http://www.flickr.com`

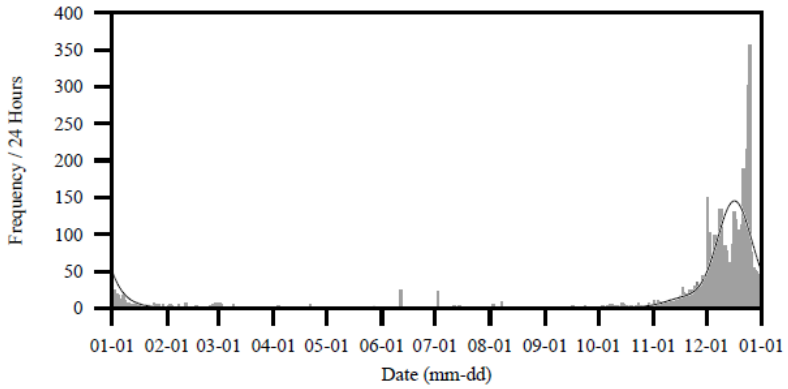


Figure 3. Distribution of "Christmas" images on Flickr (Blamey et al., 2013).

temporal expressions and temporal relations, by relying on external sources of data in order to describe specific historical events or generic periods of time (e.g. holidays), the approaches proposed are focused on specific expressions or periods of time, and they do not attempt to create a generic normalisation model to describe imprecision in temporal data among the different kinds of imprecise temporal expressions. Our work does goes further, not tackling exactly the same problem as the related work, and that it is therefore not directly comparable.

In this work, we assume that query times are grounded and known. However, this is in itself a significant task, covered in the literature (Kanhabua and Nørvåg, 2010). Knowledge base population has included a simplified version of the temporal bounding task, with maximum and minimum bounds for start and end times, and a corresponding evaluation scheme (Ji, Grishman, Dang, Griffitt and Ellis, 2010; Amigó, Artiles, Li and Ji, 2011).

3. Imprecise Temporal Data in Text

Considerable effort has been put into the extraction of temporal information from natural language texts, allowing systems to deal with complex temporal questions. However, the temporal intervals expressed in natural language are often vague, making it necessary to extend traditional temporal reasoning formalisms to cope with the vagueness (Schockaert et al., 2008). Imprecise timexes make it hard to evaluate whether events should be included in a query result that involves timeline evaluation.

Figure 4 ilustrates the importance of dealing with imprecise points in time. A query system performing searches over extracted events should be able to find those bounded by a certain period of time. Given two events e_1 and e_2, each one associated with a temporal expression t_1 and t_2, where t_1 is a precise DATE that makes it possible to place e_1 in a specific point within a timeline, and t_2 is an imprecise reference in the form "approximately N days later" which makes it impossible to know the exact day when event e_2 occurred. However, it can be reasoned the e_2 occurred after e_1. Considering a query that performs a search within the period bounded by q_b and q_e, where: $q_b < t_1 < q_e$ and $q_e < t_1 + N$, we can surely affirm that e_1 would be part of the search result. On the other

Table 2. English (En) and Portuguese (Pt) corpora analysed about the occurrence of precise and imprecise temporal expressions.

Corpus	Lang	Docs	Description
AQUAINT	En	73	News reports, also referred to as the Opinion Corpus, annotated with time expressions (Pustejovsky et al., 2010).
TE3 Platinum	En	20	The corpus used to rank participant systems in the TempEval-3 evaluation exercise, consisting of newswire documents and blog posts annotated for events, time expressions and relations (UzZaman, Llorens, Derczynski, Allen, Verhagen and Pustejovsky, 2013).
TE3 Silver	En	2,452	Documents automatically annotated as a silver standard in TempEval-3 (UzZaman et al., 2013).
TimeBank	En	183	News articles annotated with temporal information, events, times and temporal links between events and times (Pustejovsky, Hanks, Sauri, See, Gaizauskas, Setzer, Radev, Sundheim, Day, Ferro et al., 2003).
WikiWars	En	22	Documents sourced from Wikipedia, within the domain of military conflicts, containing timex annotated with TIMEX2 (Mazur and Dale, 2010).
CSTNews4	Pt	50	A discourse-annotated corpus for fostering research on single and multi-document summarization from news texts (Cardoso, Maziero, Jorge, Seno, Di Felippo, Rino, Nunes and Pardo, 2011).
* THYME	En	248	Clinical narratives datasets used in SemEval-2015 Clinical TempEval Task (Bethard et al., 2015).
* SLAM	En	1,000	Medical records without any pre annotated timexes provided by the Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London (Stewart, Soremekun, Perera, Broadbent, Callard, Denis, Hotopf, Thornicroft and Lovestone, 2009).
* InfoSaude	Pt	3,360	Medical records without any pre annotated timex extracted from the *InfoSaude* system, Public Health Department in Brazil (Bona, 2002).

* Clinical corpora

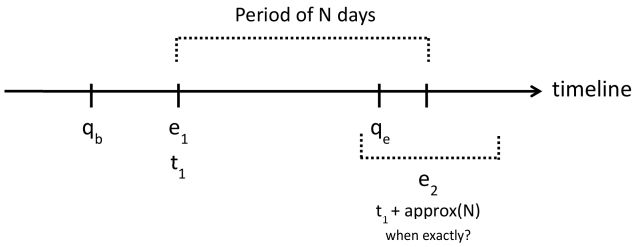


Figure 4. Example of an event (e_2) placed in an imprecise point in time.

hand, it is not possible to evaluate whether e_2 is part of the same query result, as the numerical reference that surrounds the placement of e_2 within the timeline comprises a degree of vagueness that makes it impossible to say the exact date when e_2 happened.

In this Section we show the motivation for this work by quantifying the number of imprecise temporal expressions found in different corpora. We also propose a classification for imprecise timexes.

Table 3. Occurrence of imprecise timexes in non-clinical and clinical corpora.

Non-clinical corpora				Clinical corpora			
	Total number of Timexes	Imprecise Timexes	Imprecise %		Total number of Timexes	Imprecise Timexes	Imprecise %
AQUAINT	463	35	7.6%	Thyme	3,358	659	19.6%
TE3 Platinum	158	20	12.7%	SLAM	35,120	12,226	34.8%
TE3 Silver	15,191	863	5.7%	InfoSaude	503,005	53,830	10.7%
TimeBank	478	60	12.6%	General	134,388	13,785	10.3%
WikiWars	862	112	13.0%	Gynecology	66,021	5,452	8.3%
CSTNews4	444	32	7.2%	Nutrition	64,282	6,286	9.8%
				Psychiatry	238,314	28,307	11.87%
Total (micro)	17,596	1,122	6.4%	Total (micro)	541,483	66,715	12.3%
Total (macro)			9.8%	Total (macro)			21.7%

Table 4. Occurrence of imprecise timexes by temporal granularity.

Temporal Granularity	Non-Clinical Corpora	Clinical Corpora
Year	28.5%	21.1%
Month	20.1%	21.2%
Week	7.7%	6.8%
Day	10.7%	17.6%
Time (Hour, Minute and Second)	4.9%	2.8%
Undefined	23.8%	15.2%
Others*	4.3%	15.3%

* "Others" includes Century, Decade, Quarter and Season.

3.1. Quantifying Imprecise Timexes

In order to understand the relevance of normalising imprecise temporal information in different domains, we analysed a set of three clinical and six non-clinical corpora in English and Portuguese (Table 2) to compare the occurrence of imprecise timexes in both general and specific domain data. We used the HINX system (Tissot, Gorrell, Roberts, Derczynski and Fabro, 2015) to identify the occurrence of imprecise timexes. HINX asserts a specific annotation feature (*precision* = "*imprecise*") to identify imprecise timexes, based on a set of rules to identify words, expressions and specific language structures that represent imprecision.

Table 3 compares the number of imprecise temporal expressions against the total number of timexes in each corpus, and shows that imprecise timexes in clinical corpora can reach almost 35% (SLAM corpus, 34.8%) of the temporal expressions. The percentage of imprecise expressions found in newswire was no more than 13% (WikiWars corpus).

Table 4 describes the distribution of imprecise timexes in terms of temporal granularity. The temporal granularity is the time granularity used to compose the timex, as DAY in "in less than 15 days", or UNDEFINED in "more recently". The set of expressions with granularity YEAR, MONTH, WEEK and DAY represents more than 60% of the total amount of imprecise expressions in both clinical and non-clinical corpora. Imprecise expressions denoting time (HOUR, MINUTE, and SECOND) represent less than 5% of imprecise expressions in non-clinical data and less than 3% in clinical corpora.

Table 5. Occurrence of imprecise timexes by class in clinical corpora.

	DATE			TIME			DURATION			SET		
Corpus	Tot	Imp	%	Tot	Imp	%	Tot	Imp	%	Tot	Imp	%
THYME	2,588	460	17.8%	118	13	11.0%	434	150	34.6%	218	36	16.5%
SLAM	22,678	9,296	41.0%	919	27	2.9%	8,001	2,801	35.0%	1,558	102	6.5%
SMS	210,596	19,082	9.1%	63,468	71	0.1%	190,411	34,524	18.1%	38,530	153	0.4%
General	59,835	4,838	8.1%	15,530	11	0.1%	49,829	8,900	17.9%	9,194	36	0.4%
Gynecology	33,965	1,642	4.8%	3,996	4	0.1%	24,088	3,783	15.7%	3,972	23	0.6%
Nutrition	23,324	1,969	8.4%	8,444	15	0.2%	26,933	4,285	15.9%	5,581	17	0.3%
Psychiatry	93,472	10,633	11.4%	35,498	41	0.1%	89,561	17,556	19.6%	19,783	77	0.4%
Avg (micro)	235,862	28,838	12.2%	64,505	111	0.2%	198,846	37,475	18.8%	40,306	291	0.7%
Avg (macro)			22.6%			4.7%			29.2%			7.8%

Finally, Table 5 shows the distribution of imprecise temporal expressions found in clinical corpora according to each of the main temporal classes defined by TimeML (DATE, TIME, DURATION, and SET). The occurrence of imprecise timexes is concentrated on the classes DATE and DURATION for clinical documents. In a similar analysis, we observed the occurrence of imprecise timexes is concentrated on the class DURATION in non-clinical documents.

3.2. Classification of Imprecise Timexes

We analysed the full set of imprecise expressions found in clinical corpora in order to understand the different ways the imprecision can be expressed in natural language. We defined 6 main groups of imprecise timexes according to their main language elements:

1. **Present Reference (PR)**: a time reference related to the present, based on the document creation time (DCT) (e.g. “now”, “recently”, “currently”);
2. **Modified Value (MV)**: an imprecise timex comprising a modified precise amount of time (e.g. “approximately 10 days”, “less than a month”);
3. **Imprecise Value (IV)**: an expression built around a certain imprecise amount of time (e.g. “some days”, “several weeks”), or formed with undetermined amount of time, in which granularity is usually presented in the plural, with the absence of numeric values (e.g. “years”);
4. **Range of Values (RV)**: an amount of time defined by boundaries (e.g. “every 3-4 months”, “between 8-10 years”);
5. **Partial Period (PP)**: a portion of time within a larger time frame (e.g. “the end of last year”, “middle of January”);
6. **Generic Expression (GE)**: an expression denoting a generic period or amount of time (e.g. “this time”, “at the same time”).

Table 6 details the number of imprecise timexes found in each clinical corpus according to the imprecise group. A similar distribution was also observed in non-clinical corpora. We chose to apply and test our proposed methodology starting by the three most representative kinds of imprecise expressions in terms of occurrence (PR, MV, and IV). The PR imprecise type represents more than 50% of imprecise timexes in the clinical domain. However, it comprises expressions devoid of a temporal granularity, requiring distinct questionnaire design and input data representation.

Table 6. Timexes by Imprecise Type in clinical corpora.

	Clinical Corpora		
Imprecise Type	THYME	SLAM	InfoSaude
PR	55.7%	58.0%	30.2%
MV	15.5%	6.6%	27.0%
IV	11.9%	14.4%	24.9%
RV	10.2%	4.0%	13.6%
PP	6.2%	3.2%	4.3%
GE	0.5%	13.8%	0.0%
TOTAL	659	12,229	53,830

4. Normalisation of Imprecise Timexes

Normalisation of an imprecise temporal expression depends on how people reason about imprecise information. Reasoning about an imprecise timex in a specific context, such as in clinical text, may depend on a broader narrative analysis, and an understanding of the context in which the expression was created. Despite this possible influence of different contexts on the interpretation of imprecise timexes, we present a methodology on how to produce normalisation models for each different imprecision type according to the people's common cognitive perception of temporal imprecision. Therefore, we collected and pre-processed data on how people interpret vague descriptions of time in text, and we compared different approaches in order to create and select the most appropriate normalisation model.

4.1. Specification of the Input Data

In order to collect data on how people interpret vague descriptions of time in text, we designed questionnaires[9] in two different languages (Portuguese and English). The design of the questionnaires were necessary since there is not an available dataset/standard for analysing imprecise timexes. Each question aims to capture the perception about an imprecise value for a given imprecise timex, showing a sentence comprising 2 to 3 descriptions of time that could be precise or imprecise. The target imprecise timex to be evaluated is underlined. The Portuguese questionnaire comprises 125 questions split into 5 questionnaires (25 questions each), each question made with modified (in order to guarantee de-identification) sentences found in a set of medical records from the *InfoSaude* corpus. The English version has a total of 150 questions split into 10 questionnaires (15 questions each), each question designed using fictional text to capture the perception about specific imprecise value for a given set of imprecise timexes (non-clinical).

Inter-annotator agreement (IAA) is usually used to measure the quality of a data set, by seeing how closely people agree on some objective task that is assumed to have a definitive answer, e.g. extraction of some phenomenon from text. In such a case, we would expect annotators to converge on a common value, assuming the data quality is high. Although we are asking people to fill in a questionnaire with a subjective opinion (i.e. not asking them to extract an objective

[9] `https://github.com/HeglerTissot/itn/tree/master/Questionnaire`

Table 7. Types of questions in each questionnaire and inter-annotator agreement.

Imprecise Type	Question Type	#Questions Port	#Questions Eng	#Answers (avg) Port	#Answers (avg) Eng	Fleiss' Kappa agreement Port	Fleiss' Kappa agreement Eng
MV	Approximately	30	38	70.4	88.7	0.329	0.322
	Less Than	18	26	71.3	88.8	0.285	0.324
	More Than	24	26	70.5	89.5	0.248	0.347
IV	Imprecise Value	41	48	70.2	89.4	0.198	0.201
PR	Present Reference	12	12	69.4	91.2	0.321	0.427
	Total	125	150	70.3	89.3	0.268	0.297

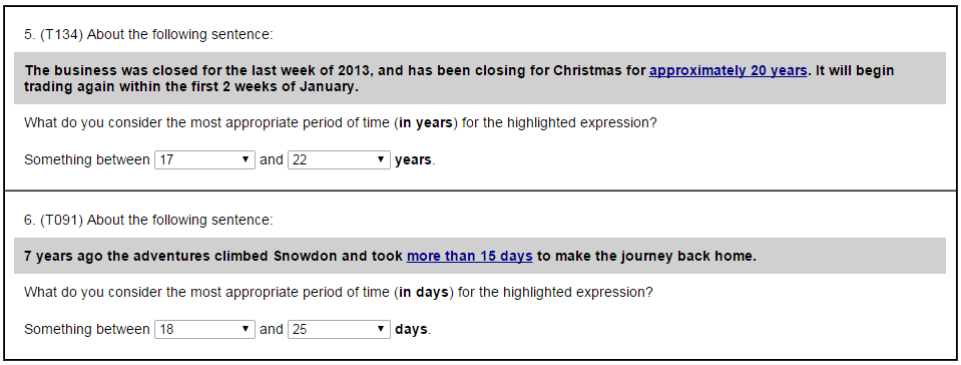


Figure 5. Example of questions used to design the questionnaire in English.

fact from the text), we used Fleiss' kappa (Fleiss et al., 1971) as a statistical measure for assessing the reliability of agreement when a fixed number of raters assign categorical ratings to a number of items. The types of questions covered by each questionnaire, average number of answers, and the inter-annotator agreement are detailed in Table 7.

MV and IV questions in the Portuguese survey asked for a specific number of days, weeks, months, or years (e.g. for "more than 10 days", one specific number of days should be selected, with options ranging from 7-60 days). The same type of question in English asked for a possible range of time (e.g. for "more than 5 days", a range of days start-end should be selected, with start point ranging from 0-40 days and end point ranging from 0-60 days). An additional option "more than 60 days" was also included on the questions covering the MV imprecise type. PR questions ("now", "currently", "recently") asked for a temporal granularity that would better describe when the associated event starts. We wanted to test different ways to answer each question, leading to the mentioned differences in the design of each questionnaire in terms of how the answers should be entered. Figure 5 shows examples of questions extracted from the questionnaires in English.

As most of the imprecise temporal expressions found in the documents we had previously analysed refer to the classes DATE and DURATION, we considered "1 day" as being the basic and minimal unit of time in the experiments. We used a discrete set of an integer number of days, disregarding granularities having TimeML TIMEX3 type TIME (hours, minutes and seconds).

The Portuguese survey was approved by the *InfoSaude* Research Committee

and submitted to 50 universities in Brazil, covering students and staff member from different departments, from which we gathered a total of 352 submissions – each question had in average 70 responses. The English survey was approved by the University of Sheffield's Research Ethics Committee and submitted to all student and staff members of an opt-out mailing list in that institution. We gathered a total of 890 submissions in English – each question had in average 90 responses.

4.2. Membership Functions

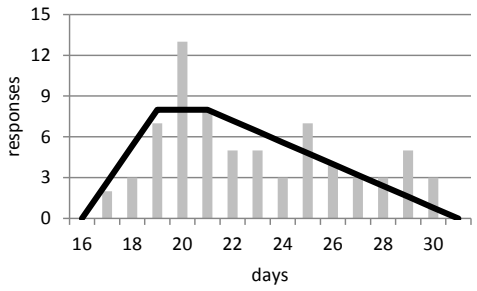
We aim to normalise imprecise expressions through the use of fuzzy membership functions (MSF). The MSF would place an imprecise timex in the timeline with a certain confidence level. In addition, a search result would have additional information indicating the confidence score for each event associated to an imprecise timex. Given a list of MSFs for the same kind of imprecise expression (e.g. of the form "less than N days"), we want to produce a generic model where, given N as an input, the model can calculate the parameters to describe a MSF for all expressions of that type.

We used two types of MSFs in our experiments: trapezoidal (4-point-based) and hexagonal (6-point-based) membership functions. Trapezoidal and hexagonal membership functions were chosen because: a) they are asymmetrical and can have their shapes adapted flexibly to match different patterns, and b) their linear boundaries make them easier to use in terms of computing fuzzy logical and relational operations.

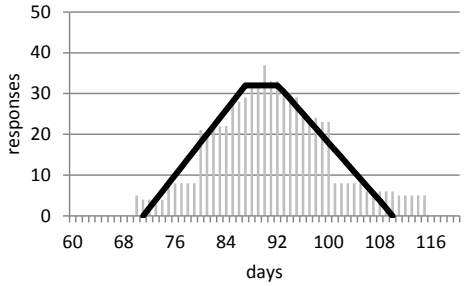
A trapezoidal MSF is defined by a set of 4 parameters (p, r, s, v), such as $M_4(x) : \mathbb{I} \rightarrow [0; 1]$, and $p < r \leq s < v$. Definition parameters p and v are the boundary limits where the confidence is 0, r and s are the boundary limits where the confidence is 1. When $r = s$, the MSF shapes like a triangular function. The MSF parameters p, r, s, v are equivalent to values a, b, c, d in Figure 1.

Similarly, a hexagonal (6-point-based) MSF is defined by a set of 6 parameters (p, q, r, s, t, v), such as $M_6(x) : \mathbb{I} \rightarrow [0; 1]$, and $p < q < r <= s < t < v$, and additionally the the trapezoidal boundaries, q and t are the values where the confidence is 0.5. In this work we refer to trapezoidal and hexagonal MSFs as by their definition parameters, using the notation $M_4(x, [p, r, s, v])$ and $M_6(x, [p, q, r, s, t, v])$.

For each question within the questionnaires we attempted to best approximate the corresponding M_4 and M_6 membership functions with respect to their definition parameters. For each question we calculated a histogram based on the number of answers given to each possible option. Then, each histogram was approximated to a trapezoidal and to a hexagonal membership function, using a full search method in order to minimise the approximation error. We looked for the best combination of values for the parameters (p, r, s, v) or (p, q, r, s, t, v), and the best MSF height in the y axis, which corresponds to the number of given answers. Figure 6(a) shows the histogram and trapezoidal function obtained for the expression "less than 30 days" from the survey in Portuguese, defined as $LessThen_{P30D}(x_{days}, [16, 19, 21, 31])$ – parameters (p, r, s, v) represent number of days, and the confidence = 1 at the height = 8 in the histogram. Similarly, Figure 6(b) presents the histogram and approximated trapezoidal function for the expression "about 3 months" from the questionnaire in English, defined as



(a) "less than 30 days" (Portuguese)



(b) "about 3 months" (English)

Figure 6. Histogram and trapezoidal MSF for two imprecise timexes.

$Approx_{P3M}(x_{days}, [71, 87, 92, 110])$ – the confidence = 1 at the height = 32 in the histogram.

4.3. Normalisation Models

We compared different approaches, such as linear regression and multilayer perceptron (Bishop, 1995), to model each kind of imprecision. In order to identify which method best models each group of imprecise timex, we explored a diverse set of alternatives. The following steps were performed to analyse the data collected from the questionnaire described in Section 4.1:

1. We started by spliting the total set of answers into two datasets (50%:50%) to be used as training and test datasets. Input data collected from the questionnaire was pre-processed. For every question we calculated the distribution of answers in the form of a histogram. A trapezoidal and a hexagonal membership functions were approximated to describe the given histogram, as described in the previous subsection.
2. For those questions using temporal granularity other than "DAY" we attempted to use both options when training the models, (a) the original granularity and the numeric value (Val) extracted from the temporal expression as it was with its original granularity (e.g. "3" in "about 3 months"), and (b)

$$
\begin{aligned}
LessThan(n) &= [0.7 * n, 1.0 * n] \\
Approx(n) &= [0.8 * n, 1.2 * n] \\
MoreThan(n) &= [1.0 * n, 1.3 * n] \\
FEW &= [2, 3] \\
SOME &= [4, 5] \\
MANY &= [6, 8] \\
SEVERAL &= [9, 12] \\
Undefined &= [8, 20]
\end{aligned}
$$

Figure 7. Unsupervised baseline parameters for IV and MV expressions.

Table 8. MLP parameters and features used.

Type	Name	Description (Value)
Features	Granularity	Four input values to set the temporal granularity – "Val" variation
	Reference Value	Number extracted for MV expressions – "Val" variation
	Reference Days	Number of days extracted for MV expressions – "Day" variation
	Temporal Context	Number of days that represents the temporal context – IV expressions
	Imprecise Value	Five input values to set the imprecise value – IV expressions
Training Parameters	maxIteration	Maximum number of training iterations to be performed (5000)
	minIteration	Minimal number of iterations to be performed before stopping (1000)
	maxNoBetter	Training stops after 200 iterations with no improvement (200)
	K	Number of folds in K-Fold Cross Validation (4)
MLP Design	hiddenLayer	Number of neurons in the hidden layer ($(inputLayerSize - 1) * (outputLayerSize - 1)$)
	outputLayer	Number of neurons in the hidden layer to produce trapezoidal MSFs (4) or trapezoidal MSFs (6)
	learningRate	Learning rate used by the backpropagation algorithm (0.95)

the same expression converted to the granularity of days (Day) (e.g. "3" in "about 3 months" was converted to "90 days").

3. For each expression type, we defined range-based unsupervised parameters to use as baseline, which were arbitrary, manually chosen. Figure 7 shows the unsupervised interval parameters defined for MV and IV questions. Each range $[b, e]$ was mapped to a $MSF(x, [b-1, b, e, e+1])$ along the experiments. For the modifier `MANY`, for example, the range value $[6, 8]$ is equivalent to a $MSF(x, [5, 6, 8, 9])$.
4. In order to produce a generic model that could be used to calculate any membership function for a given imprecise timex type, we applied four different variations of a linear regression to generalise each one of the parameters used to define trapezoidal (p, r, s, v) and hexagonal (p, q, r, s, t, v) membership functions for each given type of imprecise timex: a) the usual $(y = a + b * x)$ linear regression (Lin-A); b) we forced the independent constant a in the linear formula to be equals to zero (Lin-0); c) the linear regression with the natural logarithm values of each expression $(ln(y) = a + b * ln(x))$, in an attempt to map those expression given in terms of years (e.g. "5 years" = "1825 days") as close to those describing periods of days or weeks (Log-A); and lastly, d) the

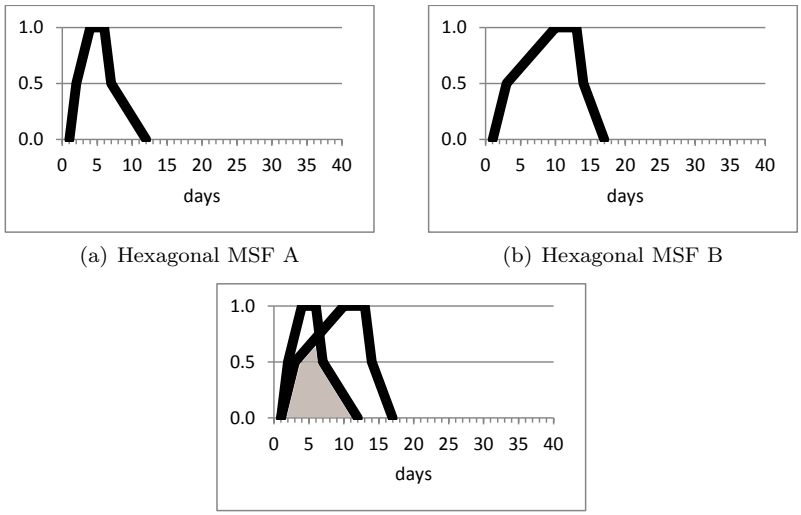
linear regression based on the logarithm values was extended to force $a = 0$ (Log-0).

5. For those timexes comprising imprecise values (IV), we also calculated the mean (MEAN) values of each membership function parameter, combining the normalised values described in 2 (Val and Day) and 4 (Lin and Log).
6. For those timexes comprising imprecise values (IV) and present references (PR), we used the temporal context as input value. We considered the "Temporal Context" as the distance in days between the current date (DCT - document creation time) and the last timex mentioned in the sentence prior to the imprecise timex being evaluated. For the designed questionnaires, DCT was defined as the date when each questionnaire was published. This approach was used in an attempt to evaluate whether the perception of a present reference imprecise timex would be influenced by the temporal context distance.
7. For MV and IV types of imprecise expression, we used a multilayer perceptron (MLP) with the Backpropagation algorithm (Gardner and Dorling, 1998) to learn how to return the membership function parameters for a given imprecise timex. We also combined the normalised values described in 2 (Val and Day) and 4 (Lin and Log). We used k-fold cross-validation to select the best model with $k = 4$. The internal MLP structure and learning parameters were chosen in a previous tuning step, after testing and comparing different configuration settings. Table 8 describes features and parameters used in the training step. In order to test the hypothesis that Present Reference (PR) expressions understanding could be influenced by the temporal context, we only tested the linear regression approach for that kind of expressions.
8. In order to evaluate each model we compared each individual membership function generated by the given model with the equivalent membership functions from the testing dataset. We used the areas of each membership functions to produce the F1-score (Equation 2), which defines how much the two functions areas overlap. Partial areas that do not overlap are considered false positive and false negative areas, and the overlap is considered as a true positive area. When $F1 = 1$ both membership functions are exactly the same, and when $F1 = 0$ there is no overlap between those given functions. The F1-score for the entire model was calculated using the average F1-score from all the membership functions used to test the model.

$$F1(A, B) = \frac{2 \times CommonArea(A, B)}{Area(A) + Area(B)} \tag{2}$$

   Figure 8 shows two hexagonal membership functions – A(x,[1,3,10,13,14,17]) and B(x,[2,3,5,7,9,14]) – and the visual representation of the F1-score between A and B, meaning the percentage of the common area relative to the total area of both functions. In the illustrated example, F1-score resulted 0.6567.
9. Finally, for each type of imprecise timex, we used the average F1-score obtained from all the different expression variations and between the trapezoidal and hexagonal membership functions in order to compare and select the most appropriate normalisation model.

The linear regression model is motivated by the hypothesis that some kinds of imprecise temporal expressions (e.g. "less than x days ago" or "in approximately x weeks" could be linearly dependent on the input amount of time x. Given this



(a) Hexagonal MSF A

(b) Hexagonal MSF B

(c) F1 score representation

Figure 8. F1-score representation between membership functions A and B – partial areas that do not overlap are considered false positive and false negative areas, and the overlap is considered as a true positive area (See Equation 2).

hypothesis, the simplest and least data-hungry tools to apply are linear regression and MLP. While SVM offers higher expressivity, it also risks making mistakes with lower amounts of data, and certainly if good results can be found through LR or MLP, this result is strong on its own. Additionally, we contrasted the linear regression results with a non-linear approach. We adopted MLP as a non-linear alternative to train normalisation models for each kind of imprecise timex, and provide comparisons there.

In order to graphically represent the normalisation models, we developed a chart format where we plot both the testing data, and the produced generalisation model. Figure 9 shows how this graphical representation works. Each known membership function produced from the input data (e.g. subfigure in the top left side represents the expression "less than 30 days") is plotted as a vertical bar, with a dark central area representing the top of the MSF, where confidence is 1 – the bottom and the top of each vertical bar represent the MSF limits where confidence is 0. The grey area in the chart's background is the normalisation model resulted for the expression type "$LessThan$". Thus, when we need to normalise an unknown expression, the normalisation model will give us the parameters that describe the corresponding MSF definition for the given expression type, by taking the limits of each dark and light grey area. For example, the selected red area at the right side represents the limits for an unknown expression "less than 90 days", which would be defined as trapezoidal MSF $LessThan_{P90D}(x_{days}, [23, 65, 85, 96])$. Other examples of known MSFs represented in the same figure as vertical bars include "less than 10 days", "less than 2 weeks", and "less than 2 months" – the figure shows 10 MSFs corresponding to the test dataset for the given type of imprecise expression.[10]

[10] There are actually two distinct MSFs corresponding to the expression "less than 30 days"

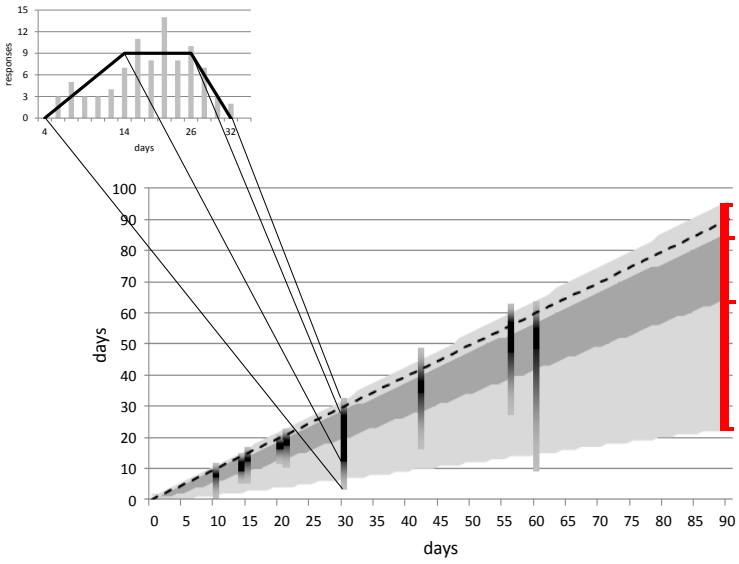


Figure 9. Graphical representation of a normalisation model.

5. Evaluation

In this section we present the results[11] of the analysis for the evaluated imprecise types (MV, IV, and PR), based on the representation model described in the previous section. We have performed a statistical hypothesis t-test for verifying the significance of the F1 scores reported for each approach. The significance threshold was set at 0.05.

5.1. Modified Value (MV) Expressions

Table 9 compares the results of each model used to produce trapezoidal (M_4) and hexagonal (M_6) membership functions for the group of expressions comprising "less than", "more than", and "approximately" subtypes for both languages (Portuguese and English). Different models are compared using the average (Avg) score between M_4 and M_6. We highlight in boldface the best Avg score for each approach (Regression and MLP) in each language (English and Portuguese).

The Log-A variation achieved the best score for this kind of expression among all the Linear Regression variations for both Languages. The MLP approach produced a result that is better than the Log-A regression variation in Portuguese. However, MLP achieved a result that is similar to the baseline in English.

in Figure 9, resulted from two different questions in the survey, but their representation is cloudy.

[11] See `https://github.com/HeglerTissot/itn` for further details about the questionnaires used in this work and the resulting models for the studied imprecise time expressions.

Table 9. F1-scores for MV temporal expressions in Portuguese and English.

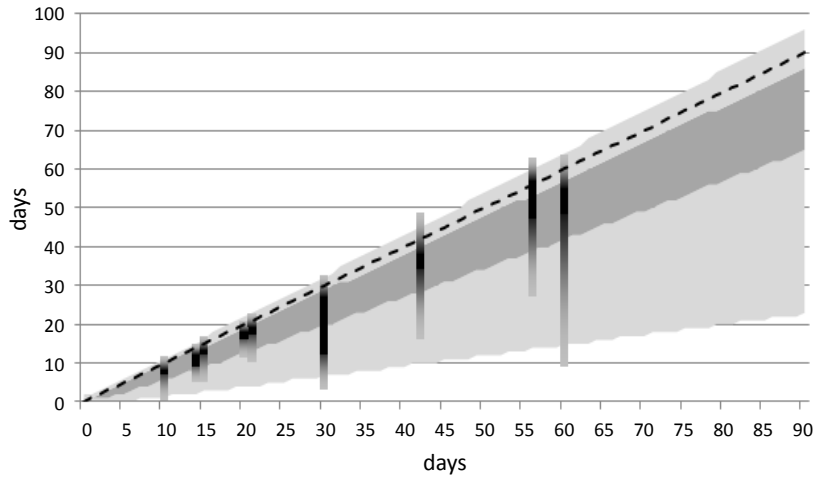
		Portuguese			English		
Method	Var	M_4	M_6	Avg	M_4	M_6	Avg
Baseline		0.673	0.646	0.660	0.741	0.731	0.736
Regression	Lin(A)	0.635	0.558	0.597	0.615	0.613	0.614
Regression	Lin(0)	0.762	0.740	0.751	0.797	0.794	0.796
Regression	Log(A)	0.772	0.746	**0.759**	0.814	0.806	**0.810**
Regression	Log(0)	0.669	0.661	0.665	0.678	0.693	0.686
MLP	Day/Lin	0.321	0.584	0.452	0.340	0.514	0.427
MLP	Day/Log	0.729	0.755	0.742	0.679	0.786	0.733
MLP	Val/Lin	0.785	0.742	**0.763**	0.738	0.787	0.763
MLP	Val/Log	0.757	0.738	0.747	0.760	0.774	**0.767**

For both languages, the t-test evidences significant differences when comparing the best MLP against the best Regression F1 scores, considering the significance threshold set at 0.05: a) in English, `p-value=0.000481` when comparing the results between Regression-Log(A) and MLP-Val/Log approaches; b) in Portuguesse, `p-value=0.003243` when comparing the results between Regression-Log(A) and MLP-Val/Lin approaches. In addition, we also compared the results between Regression-Lin(0) and Regression-Log(A), from which we found no significant differences for both languages (`p-value=0.183702` for English; `p-value=0.314776` for Portuguese. The Lin(0) variation does not rely on logarithmic transformations and this model can be directly calculated by applying simple linear transformations on the input imprecise expression.

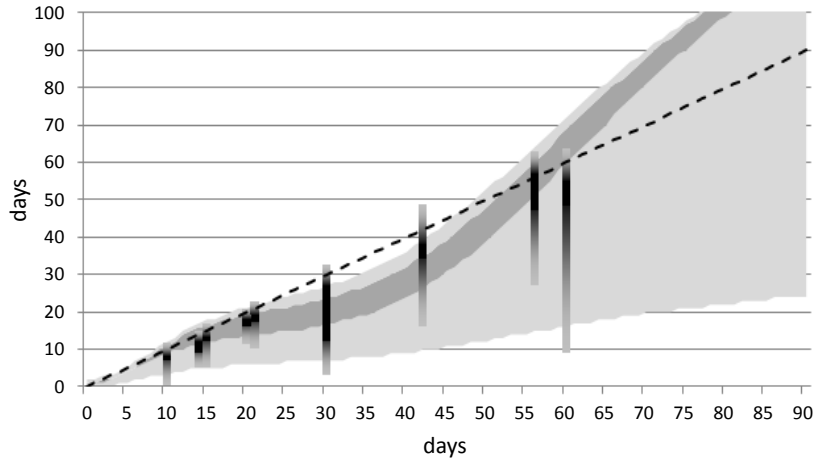
Figure 10(a) shows the model using the Log-A Linear Regression variation, and Figure 10(b) shows the model from MLP-Val/Log, both used to produce trapezoidal functions for expressions of the form "less than N days" in English. The MLP model is consistent when producing membership function parameters that are inside the limit boundaries used to train the given model. However, it is not consistent when trying to produce membership function parameters that are outside those limits. For instance, it finds values for the parameters r and s that are greater than N for "less than N days" for each $N > 60$ (darker grey area in the chart). Similar differences between Linear Regression and MLP approaches were observed in the Portuguese models. Linear Regression models are more consistent when generalising MV imprecise timexes.

Although the MLP approach resulted better for one of the languages, its inconsistency when dealing with imprecise expressions outside the limit boundaries used to train the model impose limitations and restrictions for its use. Lin-0 and Log-A models are more efficient and stable on generalising this kind of temporal imprecision, and their statistical similarity led us to believe the simplicity and stright forward applicability of the Lin-0 model makes it strongly recommended to model MV imprecise expressions. In Table 10, we present the factors $[B_p, B_r, B_s, B_v]$ used to calculate the parameters $[p, r, s, v]$ that define trapezoidal MSFs for MV temporal expressions in both languages for a given amount of time in a temporal granularity (N_{tgran}). For example, the expression "less than 30 days" in English is defined as:

$$
\begin{aligned}
LessThan_{n_{days}} &= MSF(x_{days}, [n * 0.3693, N = n * 0.7964, n * 0.9371, n * 1.0803]) \\
LessThan_{30days} &= MSF(x_{days}, [30 * 0.3693, 30 * 0.7964, 30 * 0.9371, 30 * 1.0803]) \\
&= MSF(x_{days}, [11, 23, 28, 32])
\end{aligned}
$$



(a) Linear Regression (Log-A)



(b) MLP (Val/Log)

Figure 10. Generalisation of "less than X days" expressions within the period of 0-90 for two different approaches in English.

(3)

5.2. Imprecise Value (IV) Expressions

Table 11 compares the results of each model used to produce trapezoidal and hexagonal membership functions for the IV type of temporal expressions. Linear Regression and MLP methods used the distance in days (Temporal Context) to the last precise temporal expression found in the text prior to the target im-

Table 10. Linear regression factors $[B_p, B_r, B_s, B_v]$ used to produce the parameters $[p, r, s, v]$ that define Lin-0 trapezoidal MSFs for MV expressions in Portuguese (Pt) and English (En).

Modifier	**Pt** $[B_p, B_r, B_s, B_v]$	**En** $[B_p, B_r, B_s, B_v]$
Approx(N_{tgran})	[0.7185, 0.9375, 0.9964, 1.2335]	[0.7101, 0.9325, 1.0602, 1.2965]
LessThan(N_{tgran})	[0.6921, 0.8290, 0.8554, 0.9888]	[0.3693, 0.7964, 0.9371, 1.0803]
MoreThan(N_{tgran})	[0.9705, 1.2111, 1.2605, 1.4995]	[0.8799, 1.0704, 1.2036, 1.7093]

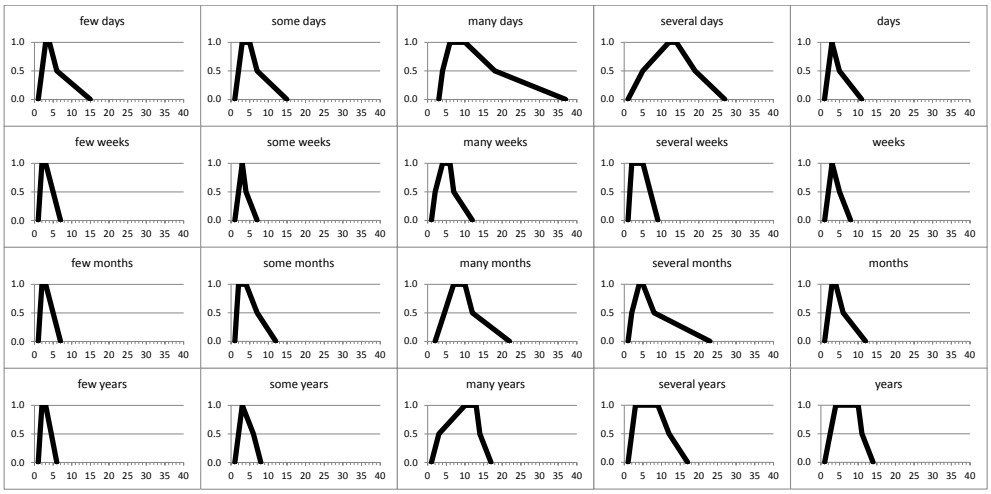
Table 11. F1-scores for IV temporal expressions in Portuguese and English.

		Portuguese			English		
Method	Var	M_4	M_6	Avg	M_4	M_6	Avg
Baseline		0.325	0.311	0.318	0.318	0.298	0.308
Mean	Day/Lin	0.661	0.847	0.754	0.866	0.847	**0.857**
	Day/Log	0.657	0.848	0.753	0.867	0.846	0.856
	Val/Lin	0.669	0.850	**0.760**	0.859	0.845	0.852
	Val/Log	0.656	0.847	0.751	0.844	0.831	0.837
Regression	Day/Lin	0.660	0.850	0.755	0.892	0.871	0.881
	Day/Log	0.660	0.846	0.753	0.884	0.868	0.876
	Val/Lin	0.673	0.858	**0.765**	0.889	0.877	**0.883**
	Val/Log	0.668	0.841	0.755	0.847	0.848	0.848
MLP	Day/Lin	0.610	0.779	0.695	0.792	0.827	0.809
(Granularity)	Day/Log	0.694	0.728	0.711	0.849	0.814	0.831
	Val/Lin	0.820	0.767	**0.793**	0.848	0.832	0.840
	Val/Log	0.751	0.726	0.738	0.848	0.819	**0.834**
MLP	Day/Lin	0.626	0.582	0.604	0.760	0.757	0.759
(Imprecise Value)	Day/Log	0.712	0.551	0.632	0.862	0.843	**0.853**
	Val/Lin	0.784	0.738	0.761	0.821	0.811	0.816
	Val/Log	0.762	0.766	**0.764**	0.841	0.762	0.802

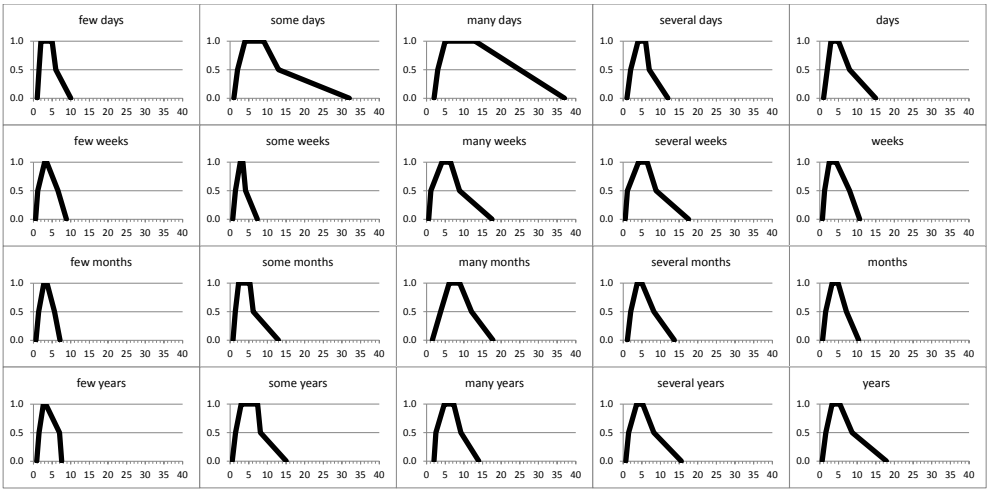
precise timex as an input parameter when creating each model. We used two MLP approaches: a) one to learn each temporal granularity ("days", "weeks", "months", "years"), and b) one to learn each imprecise value ("few", "some", "many", "several"). We highlight in boldface the best *Avg* score for each approach in each language.

The best average F1-scores for each evaluated method are similar in each language (ranging from 0.76 to 0.79 in Portuguese, and from 0.84 to 0.88 in English). The best average F1-score was achieved by the MLP model trained based on Granularities in Portuguese and by the Linear Regression (Val/Lin) in English. However, those models do not show any significant difference against the corresponding best model using the Mean method considering the significance threshold set at 0.05: a) in English, `p-value=0.075434` when comparing the Mean Day/Lin and the Regression (Val/Lin) approaches; b) in Portuguese, `p-value=0.199188` when comparing the Mean Val/Lin and the MLP Val/Lin (Granularity) approaches. The main advantage of the Mean approach refers to the fact it can be applied independently of an input value or temporal context.

Figure 11 shows the hexagonal functions created by the method Mean(Day/Lin) for the IV timexes in Portuguese and English. Table 12 presents the parameters $[p, r, s, v]$ used to define trapezoidal MSFs for IV temporal expressions in both languages. The set of parameters $[p, r, s, v]$ is given by the granularity value (Val/Lin) and by the absolute number of days (Day/Lin). For example, the expression "few weeks" in English is defined as $FewWeeks(x_{weeks}, [1, 3, 3, 9])$ (by the Val/Lin approach) or as $FewWeeks(x_{days}, [9, 20, 25, 59])$ (by the Day/Lin



(a) Portuguese



(b) English

Figure 11. Hexagonal membership functions for IV imprecise timexes comprising modifiers "few", "some", "many", and "several" combined with distinct temporal cardinalities ("days", "weeks", "months", and "years") – y-axes are the result (μ) of each MSF and x-axes represent the amount of time in the same temporal granularity corresponding to the label of each chart.

approach). Note that the approaches Val/Lin and Day/Lin produce MSFs with different temporal granularities, respectivly identified by x_{weeks} and x_{days} in each MSF definition. The former describes imprecision in the same temporal granilarity as in the original expression; the later always expresses the probabilistic distribution of a imprecise temporal expression in number of days.

Table 12. Parameters $[p, r, s, v]$ produced by the Mean approach used to define trapezoidal MSFs for IV temporal expressions in Portuguese (Pt) and English (En).

Modifier	Granularity	**Pt** Val/Lin $[p, r, s, v]$	**Pt** Day/Lin $[p, r, s, v]$	**En** Val/Lin $[p, r, s, v]$	**En** Day/Lin $[p, r, s, v]$
	Days	[1, 2, 3, 10]		[1, 3, 5, 14]	
	Weeks	[1, 3, 3, 7]	[9, 18, 22, 50]	[1, 2, 4, 11]	[7, 18, 30, 73]
	Months	[1, 2, 4, 11]	[39, 65, 110, 312]	[1, 3, 5, 10]	[26, 92, 134, 296]
	Years	[1, 3, 10, 14]	[485, 1259, 3577, 4802]	[1, 3, 4, 16]	[239, 1125, 1498, 5550]
Few	Days	[1, 2, 4, 14]		[1, 2, 4, 8]	
	Weeks	[1, 2, 3, 7]	[8, 12, 23, 50]	[1, 3, 3, 9]	[9, 20, 25, 59]
	Months	[1, 2, 3, 7]	[43, 58, 96, 198]	[1, 3, 4, 7]	[25, 80, 107, 205]
	Years	[1, 2, 3, 6]	[183, 588, 1164, 2274]	[1, 3, 4, 8]	[315, 993, 1304, 2806]
Some	Days	[1, 2, 5, 14]		[1, 3, 6, 29]	
	Weeks	[1, 3, 3, 6]	[6, 18, 23, 44]	[1, 2, 3, 6]	[6, 18, 22, 45]
	Months	[1, 2, 4, 9]	[38, 65, 129, 250]	[1, 2, 4, 11]	[27, 72, 120, 310]
	Years	[1, 2, 5, 8]	[345, 671, 1681, 2789]	[1, 3, 5, 13]	[235, 1134, 1776, 4675]
Many	Days	[3, 5, 11, 30]		[2, 5, 13, 37]	
	Weeks	[1, 3, 3, 9]	[9, 19, 23, 64]	[1, 4, 5, 15]	[6, 31, 36, 102]
	Months	[3, 6, 8, 21]	[76, 195, 254, 630]	[2, 6, 8, 17]	[59, 191, 233, 504]
	Years	[1, 10, 13, 16]	[356, 3784, 4528, 5764]	[2, 4, 7, 12]	[709, 1737, 2573, 4150]
Several	Days	[2, 8, 12, 28]		[1, 4, 5, 10]	
	Weeks	[1, 3, 3, 9]	[9, 19, 23, 64]	[1, 3, 5, 11]	[8, 24, 34, 76]
	Months	[1, 3, 5, 22]	[57, 93, 128, 663]	[1, 3, 5, 14]	[52, 81, 145, 401]
	Years	[1, 3, 10, 16]	[596, 1029, 3428, 5849]	[1, 3, 5, 14]	[261, 1280, 1583, 5081]

5.3. Present Reference (PR) Expressions

Present Reference (PR) imprecise timexes comprise those expressions including "currently", "recently", and "now". For this kind of imprecise timexes we asked people to choose the most appropriate option to express the amount of time since when the event associated with the target expression occurred. Figure 12 shows two examples of questions extracted from the English questionnaire. In each question, the target imprecise expression should be defined by another imprecise timex. Options included four IV expressions: "days", "weeks", "months", and "years".

We calculated the histogram of the given answers for each PR question, and we used the percentage of answers given to each IV expression option to create a combined membership function using a percentage of the parameters extracted from each IV expression. To calculate the linear regression model, we used the percentage of answers given for each PR question in order to produce a generic model based on the temporal context (in days). Figure 13 shows the models for two different periods (50 weeks and 20 years), including the resulted membership functions representation for each PR question in English and Portuguese. Table 13 shows the weights used to combine IV expressions with different temporal granularities in order to produce a membership function that describes each PR expression.

For example, in question number 7 (Figure 12) the expression "recently" can

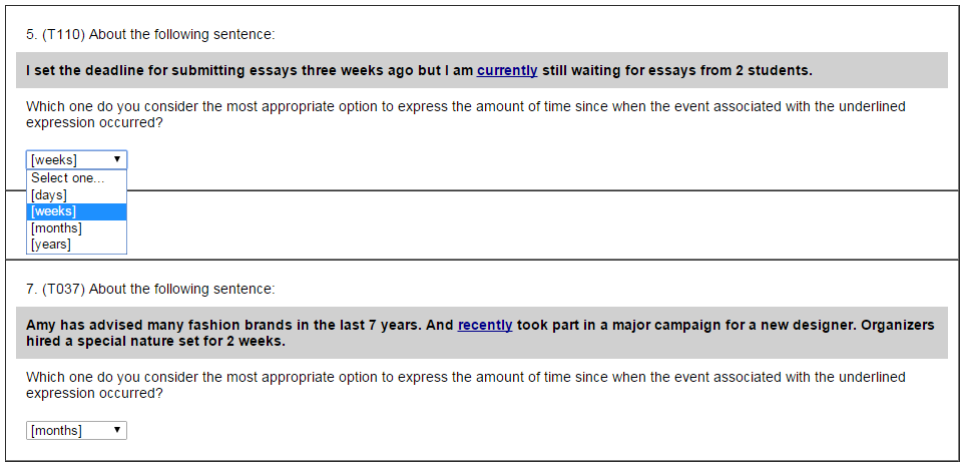


Figure 12. Example of questions covering PR imprecise timexes in English.

Table 13. Weights used to combine temporal granularities from IV membership functions to produce the parameters that define trapezoidal MSFs for PR expressions in Portuguese (Pt) and English (En).

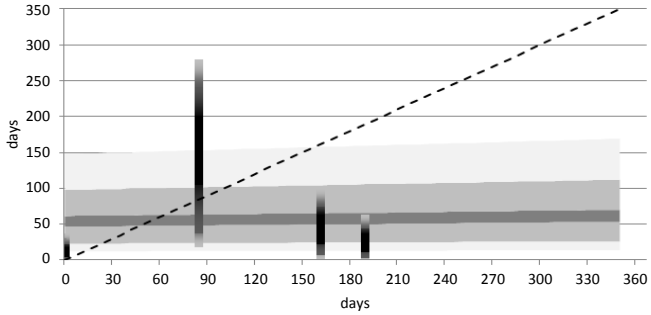
Lang	IV approach	PR expression	IV_{days}	IV_{weeks}	IV_{months}	IV_{years}
Pt	Mean	Now	0.729	0.182	0.074	0.015
		Recently	0.313	0.462	0.168	0.056
		Currently	0.379	0.308	0.231	0.081
	Regression	Now	0.784	0.130	0.075	0.011
		Recently	0.325	0.439	0.157	0.079
		Currently	0.512	0.379	0.095	0.013
En	Mean	Now	0.385	0.175	0.338	0.102
		Recently	0.180	0.528	0.281	0.010
		Currently	0.343	0.390	0.208	0.059
	Regression	Now	0.557	0.161	0.339	-0.056
		Recently	0.239	0.574	0.189	-0.003
		Currently	0.437	0.474	0.078	0.012

be mapped to a MSF by combining the IV parameters from the Mean approach:

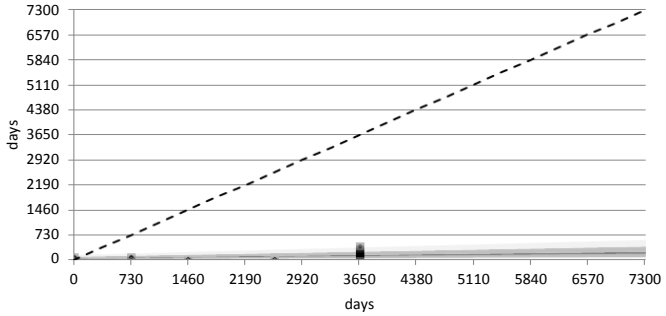
$$\begin{aligned} M_{recently} =& 0.180 \times M_{days} + \\ & 0.528 \times M_{weeks} + \\ & 0.281 \times M_{months} + \\ & 0.010 \times M_{years} \end{aligned}$$

which is equivalent to:

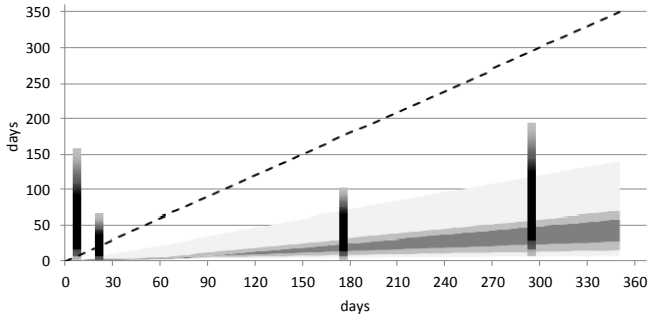
$$\begin{aligned} M_{recently} = \\ M(x, [&0.180 \times 1 + 0.528 \times 7 + 0.281 \times 26 + 0.010 \times 239, \\ &0.180 \times 3 + 0.528 \times 18 + 0.281 \times 92 + 0.010 \times 1125, \\ &0.180 \times 5 + 0.528 \times 30 + 0.281 \times 134 + 0.010 \times 1498, \\ &0.180 \times 14 + 0.528 \times 73 + 0.281 \times 296 + 0.010 \times 5550]) \end{aligned}$$



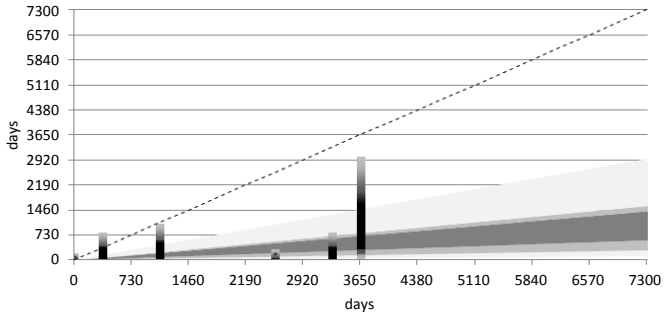
(a) 0-50 weeks in Portuguese



(b) 0-20 years in Portuguese



(c) 0-50 weeks in English



(d) 0-20 years in English

Figure 13. Hexagonal membership function model for PR imprecise timexes.

or:

$$M_{recently} = M(x_{days}, [13, 47, 70, 180])$$

PR expressions in English are more linearly dependent of the temporal context than the same expression in Portuguese. That means "recently" represents more in terms of amount of time in English when used in a temporal context of "10 years" than when it is used in a temporal context of "6 months". In the other hand, the equivalent expression in Portuguese seems to have a similar understanding independently of the temporal context being used.

The linear dependency in English and the non-linear dependency in Portuguese are confirmed by the statistical t-test when considering the significance threshold set at 0.05. We compared the PR models produced by MEAN and Linear Regression approaches from IV imprecise expressions: a) the MEAN approach is a non-temporal dependent method that uses the mean values obtained from IV expressions in order to compound PR expressions based on the average of distinct IV modifiers; b) the Linear Regression approach uses the Temporal Context as input parameter to produce MSFs. In Portuguese, Mean and Linear Regression approaches do not evidence significant differences when comparing their final scores (p-value=0.438141), whilst the same models in English present significantly different (p-value=0.015191).

We found the set of PR imprecise temporal expressions much more challengeable to model in terms of fuzzy representation. We believe further experiments focused in this specific type of imprecise temporal reference are required in order to better understand the interpretability of each possible PR expression in different contexts.

5.4. Comparing Languages

We compared models created for imprecise temporal expressions in English and Portuguese. We calculated the F1-score between both languages as the average of each F1-score calculated for each expression format for the trapezoidal and hexagonal MSFs. All expressions within the same type were combined to calculate a partial F1-score (e.g. "some days" in Portuguese and the same expression in English) as the average between F1-score for the trapezoidal and hexagonal MSFs. The calculated average F1 score among all the expressions resulted the similarity between Portuguese and English.

However, when calculating the F1-score using the MSF area, it was not possible to identify whether the differences are more concentrated in the top (confidence=1) or the bottom (confidence=0) of such functions. In order to identify how relevant such differences are, we used a variation of F1-score that we called $F1_{3D}$. We considered each MSF as a tridimensional object, from which the third dimension identifies how deep each MSF is, varying from 0 at the bottom to 1 at the top. Instead of using the MSF areas, we then used the MSF volumes to calculate $F1_{3D}$ (Equation 4). Figure 14 illustrates the difference between F1 and $F1_{3D}$, comparing three MSFs (A, B, and C). A and B have a difference in the top, whilst A and C have the exact same difference in terms of area, in the bottom instead. Thus, $F1(A, B) = F1(A, C) = 0.9655$. When calculating the $F1_{3D}$, we can observe $F1_{3D}(A, B) < F1_{3D}(A, C)$, what means A and B have differences more concentrated in the top comparatively to the differences between A and C – differences at the top have more influence to decrease $F1_{3D}$ than differences at

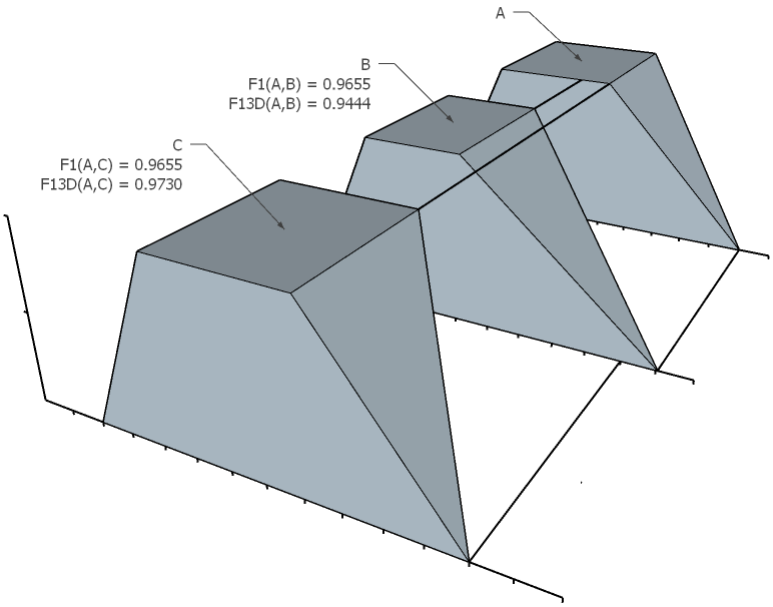


Figure 14. Contrasting F1 and $F1_{3D}$ scores used to calculate the similarity between membership functions.

Table 14. F1 and $F1_{3D}$-scores between Portuguese and English.

Imprecise Type	F1	$F1_{3D}$
MV	0.731	0.692
IV	0.767	0.719
PR	0.391	0.304

the bottom due to the MSF depth.

$$F1_{3D}(A,B) = \frac{2 \times CommonVolume(A,B)}{Volume(A) + Volume(B)} \tag{4}$$

We used the following normalisation models to compare the results in English and Portuguese: a) Log(A) regression models to compare MV expressions; b) MEAN models to compare IV expressions; and c) Lin(A) regression models to compare PR expressions. Table 14 shows the F1 and $F1_{3D}$-scores between English and Portuguese. We can observe $F1 > F1_{3D}$ for all the three types of imprecise temporal expressions analysed, indicating that differences tend to be concentrated more closely to the top of the MSFs, where the confidence is higher, and differences can be considered more relevant.

6. Conclusions

We have presented an analysis of previously unstudied imprecise time expressions (timexes) in text. This analysis helps to address the overall problem of dealing with temporal expressions in information extraction. Our work introduces

three novel techniques for this analysis. First, we provide a novel classification of imprecise timexes. Second, we develop a novel methodology to obtain membership functions for timexes, based on human interpretation of imprecise timexes. Third, as well as the usual F1-score for evaluation, we introduce a novel metric for identifying the differences between membership functions, along 3 dimensions - the $F1_{3D}$. Our models were applied to both English, and for the first time, to Portuguese expressions.

The resulting models give an insight in to the way in which imprecise expressions are interpreted in different languages. For example, the Linear Regression Log(A) membership function that defines the expression "less than 90 days" in Portuguese includes possible interpretations - albeit at a low level of confidence - of 91 to 95 days. This leads us to believe that temporal imprecision is not mathematically reasoned, and that there is a level of uncertainty that is able to cross the boundary limits defined by the numerical values found within the temporal expressions.

In future work, we plan to perform experiments to obtain normalisation models corresponding to the other types of imprecise expressions (PP, RV, and GE), and examine whether the differences between languages can be influenced by the knowledge domain or by cultural differences. We also plan further examine the relation between the F1 and $F1_{3D}$ scores and compare their interpretability against other probability distribution divergence metrics, such as the the Kullback–Leibler (KL) divergence. Additionally, we plan to compare the membership function models against other probabilistic representations (e.g. gaussian or gamma distributions), and validate in what extent such probabilistic generalisations are able to mimic the results we found in this work.

Up to 35% of temporal expressions may be imprecise in some domains. By normalising these imprecise expressions, we can greatly increase the amount of extracted events connected to a timeline. We plan to perform search-based experiments over the extracted events from medical records, in order to provide an extrinsic evaluation of the impact of dealing with such imprecise temporal data on the overall IE process.

Acknowledgments

We would like to thank the Mayo Clinic for permission to use the THYME corpus, and CAPES,[12] which is partially financing this work. This work also received funding from the European Union's Seventh Framework Programme (grant No. 611233,PHEME). AR and LD are part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London.

References

Ahn, D., Adafre, S. F. and Rijke, M. (2005), 'Extracting temporal information from open domain text: A comparative exploration', *Journal of Digital Information Management* **3**, 2005.

[12] `http://www.iie.org/en/programs/capes`

Allen, J. F. (1983), 'Maintaining knowledge about temporal intervals', *Commun. ACM* **26**(11), 832–843.

Amigó, E., Artiles, J., Li, Q. and Ji, H. (2011), An evaluation framework for aggregated temporal information extraction, *in* 'SIGIR-2011 Workshop on Entity-Oriented Search'.

Bartak, R., Morris, R. and Venable, K. (2013), *An Introduction to Constraint-Based Temporal Reasoning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool.

Bethard, S. (2013), Cleartk-timeml: A minimalist approach to tempeval 2013, *in* 'Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)', Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 10–14.

Bethard, S., Derczynski, L., Pustejovsky, J. and Verhagen, M. (2015), SemEval-2015 Task 6: Clinical TempEval, *in* 'Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)', Association for Computational Linguistics.

Bethard, S., Martin, J. H. and Klingenstein, S. (2007), Timelines from text: Identification of syntactic temporal relations., *in* 'ICSC', IEEE Computer Society, pp. 11–18.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J. and Verhagen, M. (2016), Semeval-2016 task 12: Clinical tempeval, *in* 'Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)', Association for Computational Linguistics, San Diego, California, pp. 1052–1062.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, NY, USA.

Blamey, B., Crick, T. and Oatley, G. (2013), 'the first day of summer': Parsing temporal expressions with distributed semantics, *in* M. Bramer and M. Petridis, eds, 'Research and Development in Intelligent Systems XXX', Springer International Publishing, pp. 389–402.

Bona, C. (2002), Avaliação de processos de software: Um estudo de caso em xp e iconix, Master's thesis, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina (UFSC).

Burman, A., Jayapal, A., Kannan, S., Kavilikatta, M., Alhelbawy, A., Derczynski, L. and Gaizauskas, R. (2011), USFD at KBP 2011: Entity linking, slot filling and temporal bounding, *in* 'Proceedings of the Text Analysis Conference'.

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. R., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V. and Pardo, T. A. S. (2011), Cstnews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese, *in* 'Proceedings of the 3rd RST Brazilian Meeting', Cuiabá, Brazil, pp. 88–105.

Caselli, T. (2009), Time, Events and Temporal Relations: an Empirical Model for Temporal Processing of Italian Texts, PhD thesis, Università di Pisa, Pisa, Italy.

Chambers, N. (2013), Navytime: Event and time ordering from raw text, *in* 'Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)', Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 73–77.

Coelho, A. L. V. and Raposo, A. B. (2005), Dealing with imprecision in temporal interdependencies between collaborative tasks: A fuzzy perspective, *in* 'II Workshop Brasileiro de Tecnologias para Colaboração (WCSCW 2005)', Vol. 2, Atlanta, Georgia, USA, pp. 791–800.

Davis, J. and Goadrich, M. (2006), The relationship between precision-recall and roc curves, *in* 'Proceedings of the 23rd international conference on Machine learning', ICML '06, ACM, New York, NY, USA, pp. 233–240.

Fagerberg, A. (2014), 'Temporal information extraction using regular expressions', `http://www.antonfagerberg.com/files/tempex_anton_fagerberg.pdf`. Accessed: Mar, 2014.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G. (2005), TIDES 2005 standard for the annotation of temporal expressions, Technical report, The MITRE Corporation.

Filannino, M. and Nenadic, G. (2014), Mining temporal footprints from Wikipedia, *in* 'Proceedings of the First AHA! workshop', pp. 7–13.

Fleiss, J. et al. (1971), 'Measuring nominal scale agreement among many raters', *Psychological Bulletin* **76**(5), 378–382.

Gardner, M. W. and Dorling, S. R. (1998), 'Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences', *Atmospheric Environment* **32**(14-15), 2627–2636.

ISO (2007), Language resource management — semantic annotation framework (semaf) — part 1: Time and events, *in* 'ISO/TC37/SC 4N269 rev04', ISO Report.

Ji, H., Grishman, R., Dang, H. T., Griffitt, K. and Ellis, J. (2010), Overview of the tac 2010 knowledge base population track, *in* 'Third Text Analysis Conference (TAC 2010)'.

Kanhabua, N. and Nørvåg, K. (2010), Determining time of queries for re-ranking search results, *in* 'Research and Advanced Technology for Digital Libraries', Springer, pp. 261–272.

Kolomiyets, O. (2012), Algorithms for Temporal Information Processing of Text and their Applications, PhD thesis, Informatics Section, Department of Computer Science, Faculty of Engineering Science. Moens, Marie-Francine and De Schreye, Daniel (supervisors).

Kolomiyets, O. and Moens, M.-F. (2010), KUL: recognition and normalization of temporal expressions, *in* 'Proceedings of SemEval-2 5th Workshop on Semantic Evaluation - ACL SigLex,', ACL, pp. 325–328.

Kolomiyets, O. and Moens, M.-F. (2013), KUL: A data-driven approach to temporal parsing of documents, *in* 'Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),', ACL, pp. 83–87.

Ling, X. and Weld, D. S. (2010), Temporal information extraction., *in* M. Fox and D. Poole, eds, 'AAAI', AAAI Press.

Llorens, H., Derczynski, L., Gaizauskas, R. J. and Saquete, E. (2012), TIMEN: An open temporal expression normalisation resource., *in* 'LREC', ELRA, pp. 3044–3051.

Mazur, P. and Dale, R. (2010), Wikiwars: A new corpus for research on temporal expressions, *in* 'Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing', EMNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 913–922.

Nagypál, G. and Motik, B. (2003), A fuzzy model for representing uncertain, subjective and vague temporal knowledge in ontologies, *in* 'In Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics, (ODBASE), volume 2888 of LNCS', Springer, pp. 906–923.

Pedrycz, W. and Gomide, F. (1998), *An Introduction to Fuzzy Sets: Analysis and Design*, Complex adaptive systems, NetLibrary, Incorporated.

Pustejovsky, J., Castano, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A. and Katz, G. (2003), TimeML: Robust specification of event and temporal expressions in text, *in* 'in Fifth International Workshop on Computational Semantics (IWCS-5)'.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L. et al. (2003), The TimeBank corpus, *in* 'Proceedings of the Corpus Linguistics Conference', Vol. 2003, p. 40.

Pustejovsky, J., Lee, K., Bunt, H. and Romary, L. (2010), ISO-TimeML: An international standard for semantic annotation, *in* 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', ELRA.

Sauri, R., Littman, J., Gaizauskas, R., Setzer, A. and Pustejovsky, J. (2006), 'TimeML Annotation Guidelines, Version 1.2.1'.

Schockaert, S. (2005), Construction of membership functions for fuzzy time periods, *in* 'Proceedings of the ESSLLI 2005 Student Session'.

Schockaert, S., Cock, M. D. and Kerre, E. E. (2008), 'Fuzzifying Allen's temporal interval relations.', *IEEE T. Fuzzy Systems* **16**(2), 517–533.

Stewart, R., Soremekun, M., Perera, G., Broadbent, M., Callard, F., Denis, M., Hotopf, M., Thornicroft, G. and Lovestone, S. (2009), 'The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data', *BMC Psychiatry* **9**, 51.

Strötgen, J., Zell, J. and Gertz, M. (2013), Heideltime: Tuning english and developing spanish resources for tempeval-3, *in* '2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7h International Workshop on Semantic Evaluation (SemEval 2013)', ACL, Atlanta, Georgia, USA, pp. 15–19.

Styler, W., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P., Erickson, B., Miller, T., Lin, C., Savova, G. and Pustejovsky, J. (2014), 'Temporal annotation in the clinical domain', *Transactions of the Association for Computational Linguistics* **2**, 143–154.

Sun, W., Rumshisky, A. and Uzuner, O. (2013), 'Evaluating temporal relations in clinical text: 2012 i2b2 Challenge', *J Am Med Inform Assoc* **20**(5), 806–813.

Tissot, H., Gorrell, G., Roberts, A., Derczynski, L. and Fabro, M. D. D. (2015), UFPRSheffield: Contrasting rule-based and support vector machine approaches to time expression identification in clinical tempeval, *in* 'Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)', Association for Computational Linguistics, Denver, Colorado, pp. 835–839.

UzZaman, N. and Allen, J. F. (2010), Trips and trios system for tempeval-2: Extracting temporal information from text, *in* 'Proceedings of the 5th International Workshop on Semantic Evaluation', SemEval '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 276–283.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M. and Pustejovsky, J. (2013), SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations, *in* 'Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)', ACL, pp. 1–9.

Zadeh, L. A. (1994), 'Fuzzy logic, neural networks, and soft computing', *Commun. ACM* **37**(3), 77–84.

Zhou, X., Li, H., Lu, X. and Duan, H. (2011), Temporal expression recognition and temporal relationship extraction from chinese narrative medical records, *in* 'International Conference on Bioinformatics and Biomedical Engineering', pp. 1–4.