Good Night at 4 pm?! Time Expressions in Different Cultures

Anonymous ACL submission

Abstract

We propose the task of culture-specific time expression grounding, i.e. mapping from expressions such as "morning" in English or "manhã" in Portuguese to specific hours in the day. We propose 3 language-agnostic methods, one of which achieves promising results on gold standard annotations that we collected for a small number of languages. We then apply this method to 28 languages and analyze the similarities across languages in the grounding of time expressions.

1 Introduction

Natural language understanding requires the ability to map language such as color descriptions (McMahan and Stone, 2015), spatial instructions (Chen et al., 2019), and gradable adjectives (Shivade et al., 2016) to real-world physical properties. This paper focuses on temporal grounding, particularly mapping time expressions such as "morning" and "evening" to hours in the day. Temporal commonsense reasoning has been gaining traction lately (Zhou et al., 2019; Qin et al., 2021), and this important capability can benefit various temporal tasks such as event ordering and duration prediction.

One of the challenges in grounding time expressions to standard times is that such expressions may be interpreted with some variation by different people. Reiter and Sripada (2002) found that human-written weather forecasts exhibited significant individual differences between forecasters in the interpretation of time expressions. One factor for this variation is cultural differences. Vilares and Gómez-Rodríguez (2018) analyzed the time of day in which people from 53 countries posted time-specific greetings such as "good morning" and "good evening" on Twitter. They showed variation in greeting times across languages and cultures, which they connected to known facts and published statistics about cultural differences, such as differences in average wake and sleep times.

We propose to re-frame the research question posed by Vilares and Gómez-Rodríguez (2018) as a task of time expression grounding: given a time expression, the goal is to map it to a specific range of hours during the day. For example, what is the range of times referred to by an Italian speaker mentioning pomeriggio (afternoon)? We collected gold standard interpretations from four countries, which indeed exhibited some variation. We then proposed 3 language-agnostic methods based on either a corpus or a language model (LM).

The corpus-based method performed well across languages, outperforming the method proposed by Vilares and Gómez-Rodríguez (2018) on 3 out of 4 languages. Encouraged by the performance on the labelled languages, we applied the method to additional 24 unlabelled languages, and analyzed the differences predicted by the models.

In the future, we plan to incorporate this method into NLP systems that may benefit from temporal grounding. Areas of future work involve testing our methods on low-resource languages, as well as researching ways to overcome reporting bias (Gordon and Van Durme, 2013): the under-representation of trivial facts in written text. We hope this work would be another small step in the long-term goal of developing culturally-aware commonsense reasoning models (Acharya et al., 2021).[1]

2 Data

We collected gold standard annotations for the start and end times of five time expressions: morning, noon, afternoon, evening, and night. The annotations were collected in Amazon Mechanical Turk (AMT) for English, Hindi, Italian, and Portuguese. We describe the rationale behind the choice of languages (§2.1), the annotation guidelines (§2.2), and the observations from the collected data (§2.3).

[1]Our data and code are available at https://anonymous.4open.science/r/time_expressions-23F6.

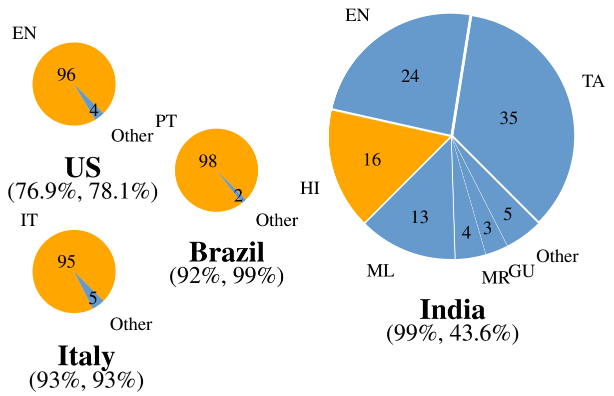

Figure 1: Percents of native languages collected from each country. India is the only country where the majority native language differs from the language used in Wikipedia and BERT (Hindi). Numbers in brackets: (1) percents of native speakers of the target language (in orange) living in this country (López, 2015); and (2) percents of the country's population that speaks this language at home (from Wikipedia).

2.1 Choice of Languages

The languages in this paper are not meant to be a representative sample of all languages. We selected these languages based on the following criteria.

Availability of AMT Workers. By and large, AMT does not facilitate filtering workers by the languages in which they are fluent.[2] We thus treated country as a proxy for language, e.g. assuming that most workers in Brazil speak Portuguese, while asking workers about their native language. AMT is available at select countries, and the number of workers in each country varies. We got the most responses from US and India (100 each), in line with published analyses of demographics (Difallah et al., 2018) and language demographics in AMT (Pavlick et al., 2014). We collected 91 responses from Brazil and 58 from Italy.

The Interplay between Country and Language. We focused on pairs of country and language where most of the country's population speaks that language, and most of the L1 speakers of the language reside in that country. For instance, 78.1% of US residents speak English at home, and 76.9% of L1 English speakers reside in the US.[3] Figure 1 shows that for 3 out of the 4 countries, the majority of workers indicated they were native speakers of the majority language. The exception is India, which has many languages. Hindi is the most spoken language in India (followed by Bengali: 8% and Telugu: 6.7%) and has the larger Wikipedia corpus and a BERT model. Among the workers from India, 16% indicated they were Hindi speakers.

While the gold standard annotations are limited to 4 languages, the framework we describe in Section 3 is unsupervised and almost entirely language-agnostic. As we discuss in Section 4.3, we applied the model to additional 24 languages, selected based on the availability of a Wikipedia corpus and an LM for that language.

2.2 Annotation Task

We asked workers to identify their native language, and posed them the following questions regarding each time expression (e.g. *noon*).

1. **What is the equivalent word for *noon* in your native language?** We allowed workers to check "There is no equivalent expression in my language".

2. **What is the range of time you consider as *noon*?** Workers were required to indicate the start and end times.

We followed with an option to add a time expression in their language that wasn't mentioned in the HIT as well as free text comments. To ensure the quality of annotations, we required that workers had a 95% approval rate for at least 100 prior HITs.

2.3 Observations

Figure 4 displays the average start and end time for each country and each time expression. Notably, morning is quite consistent across the different countries and noon is the short period around 12 pm. The variation is higher for afternoon and evening. Many workers from Brazil noted that Portuguese uses the same word for evening and night (*noite*), and that evening turns quickly into night because of the country's tropical climate. This results in a very early night time in the annotations (3:16 pm), and high overlap between the afternoon, evening, and night spans.

Workers across countries suggested a missing expression that spans the time between midnight and sunrise, which they referred to as "midnight", "after midnight", "late night", "early morning", and

[2] There is a recent qualification type for a few languages, such as Chinese and German. It is an expensive filter at an additional $1 fee per HIT. We tried collecting annotations for Chinese in German but got very few responses, likely due to the small number of workers that have these qualifications.

[3] Followed by the UK (17.6%), Nigeria (11.05%), Canada (6%), Australia (5%), South Africa (1.47%), Ireland (1.22%), and New Zealand (1.1%).

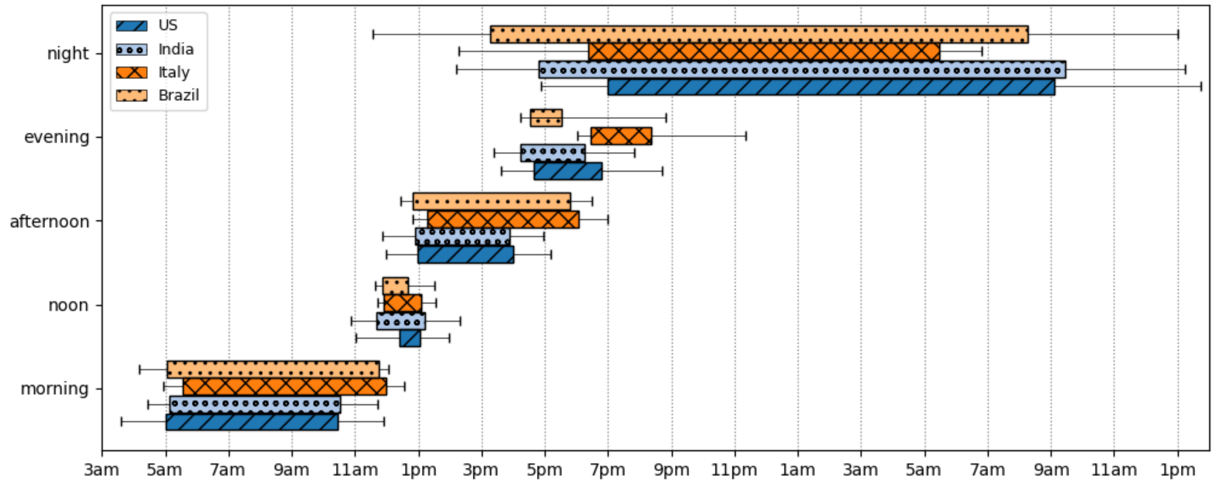

Figure 2: Start and end time distributions for each time expressions, as indicated by workers from 4 countries.

"dawn". Other suggestions included "twilight" (6-7 pm, India), "sunrise" (5-6 am, Italy), "late morning" (11-11:59 am, Italy), "after lunch" (1:15-2 pm, Italy), and "late afternoon" (3-4 pm, Italy).

Finally, some workers commented that the interpretations of time expressions varies in different seasons because of the changes in sunrise and sunset times. The data was collected in October, and although we don't know the exact location of the workers, we can test the night start and end times against the average October sunrise and sunset times in the capital of each country. Setting aside Brazil that doesn't distinguish evening and night, there is somewhat of a match between the average **sunset** time and the average *night start* time: US: **6:30 pm**/*6:59 pm*, India: **5:52 pm**/*4:49 pm*, and Italy **6:30 pm**/*6:22 pm*. There was no such match between sunrise time and the end of the night or beginning of the morning.

3 Methods

We define the time expression grounding task: given a time expression, the goal is to predict its start and end times. We developed 3 methods that differ along two dimensions: (1) the source from which the times are learned: a corpus (§3.1) or a language model (§3.2); and (2) whether to compute start and end times directly or indirectly through estimating a distribution of times.

3.1 Extractive Approach

Estimating Hour Distributions. We search Wikipedia for occurrences of a regular expression that matches a broad range of time formats, including both 24-hour and 12-hour clock formats. For each time expression X_i, we compute D_i, the distribution of hours from co-occurring time mentions within the same paragraph. For example, given the sentence "See you in the evening, at 19:30" we extract a co-occurrence of "evening" with 7 pm. We used Google Translate to translate the English time expressions to other languages.

Inferring Start and End. To infer the start and end times S_i and E_i from D_i, we define an optimization problem and formulate it as an integer linear programming (ILP) problem detailed below.

Input:
D_1 ... D_5: hour distribution per expression
Define: `// start and end variables`
(S_1, E_1) ... (S_5, E_5), $0 \leq S_i, E_i \leq 23$
Maximize:
$\sum_i \sum_h \text{WithinRange}(h, S_i, E_i) \cdot D_i[h]$
Constrained to:
`// start before end except at night`
$\forall_{i=1,\ldots,4} S_i < E_i, \;\; S_5 < E_5 + 24$
`// sort expressions`
$\forall_{i=1,\ldots,4} S_{i+1} \geq E_i$

The goal is to find a global solution for all the time expressions, with non-overlapping time ranges in which the expressions are sorted, e.g. morning comes before noon. We maximize the number of observations in D_i that are within the inferred start and end times.[4]

3.2 LM-Based Approach

We used multilingual BERT (mBERT; Devlin et al., 2019), a single BERT model trained on Wikipedia

[4]We also tried to extract start and end times directly from the corpus, but the signal was too sparse.

Template
It was [MASK] in the <time_exp> .
It is [MASK] in the <time_exp> .
It happened yesterday in the <time_exp> , at [MASK] .
It happened in the <time_exp> , at [MASK] .
It will happen in the <time_exp> , at [MASK] .
Every <time_exp> at [MASK] .
The <time_exp> starts at [MASK] .
The <time_exp> ends at [MASK] .

Table 1: Templates used by the LM-based method to predict the distribution (top) or start/end times (bottom).

in multiple languages that achieves strong zero-shot cross-lingual transfer performance (Wu and Dredze, 2019).

Method 1: Estimating Hour Distributions. For each time expression, we query BERT for substitutes for the masked token in each template in the top part of Table 1. We translated the templates to other languages using Google Translate.[5]

Since LM predictions are sensitive to the prompt, we follow Jiang et al. (2020) and aggregate the predictions across these various templates. We also allow for various time formats. For example, we query BERT for the substitutes of each of "It is [MASK]:00 in the morning", "It is [MASK].00 in the morning", and "It is [MASK] in the morning". We sum the distributions and normalize the scores for all numbers within the range of 0 and 23.

For languages spoken mostly in countries where 12-hour clock is the norm, we computed the distribution for hours in the range of 0 and 12.[6] We then assigned each hour back into the template and predicted whether the next token is more likely to be am or pm (or its equivalent in the target language). For example, if BERT assigned 9:00 a score of 0.3 in the morning distribution, and the query "It is 9:00 [MASK] in the morning" predicted am with a score of 0.9 and pm with 0.1, then in the final 24-hour clock distribution, 9 has a score of $0.3 \cdot 0.9 = 0.27$ and 21 has a score of $0.3 \cdot 0.1 = 0.03$.

Finally, we use the same ILP formulation to infer the start and end times from the hour distributions.

Method 2: Directly Predict Start and End Times. For each time expression, we separately query BERT for the substitutes of the masked tokens in the start template and end template in the bottom part of Table 1. We apply the same processing as described above. The output of this step is a start time distribution SD_i and an end time distribution ED_i over 24 hours for each time expression X_i. We infer the start and end times with the same optimization problem, but with a slightly modified objective detailed below. The objective is to select the most highly scored start and end time for each expression, that adhere to the same constraints.

Maximize:
$$\sum_i \sum_h (\mathbb{1}(S_i == h) \cdot SD_i[h] + \mathbb{1}(E_i == h) \cdot ED_i[h])$$

4 Experiments

4.1 Baseline

Our baseline is based on the Greetings method proposed by Vilares and Gómez-Rodríguez (2018). Their study focused on 4 out of the 5 time expressions used in our paper: morning, afternoon, evening, and night. We use their dataset and induce the corresponding time expression distributions. We focus on tweets in English from the US (1.34M), Portuguese from Brazil (2M), Italian from Italy (4,821), and Hindi from India (6,069). We then infer the start and end times using the ILP problem in Section 3.1. Although the dataset does not include statistics for "noon" (due to the lack of a corresponding greeting), the global objective in the ILP formulation is expected to infer the start and end times for noon based on the surrounding time expressions.

4.2 Results

Figure 3 displays the predicted start and end times for each expression according to each method, in comparison to the gold standard times of each language. For quantitative evaluation, we define minute-level accuracy. We classify each minute of the day to a time expression based on the start and end times, and compute the accuracy compared to the gold standard minute classification. Table 2 shows the accuracy as well as the average differences in hours between the predicted and gold standard start (ΔStart) and end (ΔEnd) times.

There is a general preference for the extractive method, that achieves between 65% and 90% accuracy across languages. The exception is Portuguese, where this method performs worse than the others, and in particular by the LM Start-End method that performs remarkably well. The two LM-based methods perform substantially worse on

[5] For better translation quality, we assigned "morning" into the <time_exp> placeholder and "9:00" into the [MASK] placeholder.

[6] In this paper, such languages are English and Hindi.

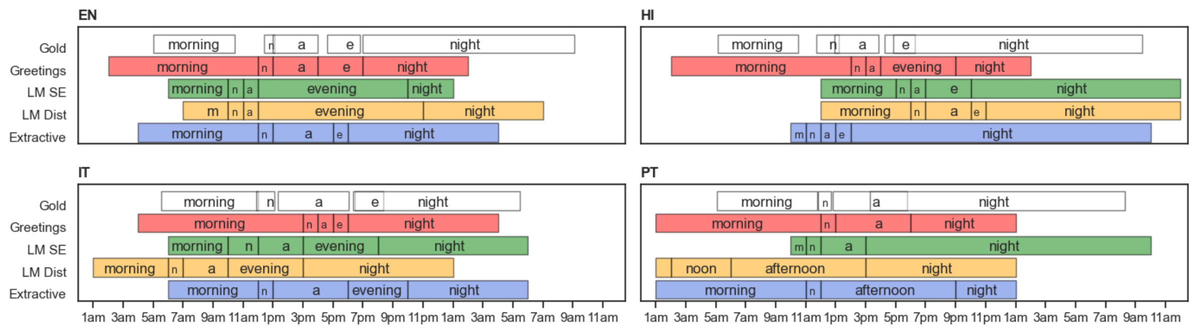

Figure 3: Start and end times for each time expressions, in English, Hindi, Italian, and Portuguese, as estimated by each method and compared to the gold standard.

Model	Type	Acc.	ΔStart	ΔEnd
		EN		
Extractive	**Distribution**	**84.3**	**0.6**	**1.7**
LM	**Distribution**	63.3	3.0	2.6
	Start-End	49.2	2.6	3.6
Greetings	**Distribution**	80.7	0.8	1.8
		HI		
Extractive	**Distribution**	**80.4**	2.5	**1.9**
LM	**Distribution**	54.2	6.2	5.3
	Start-End	58.4	5.0	4.1
Greetings	**Distribution**	60.7	**2.4**	3.1
		IT		
Extractive	**Distribution**	**90.1**	**1.0**	**0.5**
LM	**Distribution**	55.3	5.7	6.0
	Start-End	80.3	1.7	1.4
Greetings	**Distribution**	71.9	1.8	2.2
		PT		
Extractive	**Distribution**	65.0	2.9	3.0
LM	**Distribution**	77.3	5.2	6.6
	Start-End	**95.5**	**1.7**	**1.5**
Greetings	**Distribution**	79.5	4.7	4.7

Table 2: Minute-level accuracy and differences in gold and predicted start and end times across languages.

the other languages. Finally, the results for India are surprisingly not bad despite the mismatch between the native languages of the annotators and the language used by our methods.

4.3 Application to Other Languages

We applied our proposed methods to additional unlabelled languages detailed in Table 3. The languages are sorted according to their Wikipedia corpus size. The Table shows the predicted start and end time for each language and each time expression. Without labelled data it is hard to judge the correctness of the predictions, but the predictions of some languages seem more reasonable than others. In particular, we observed that some time expressions appeared in the corpus more frequently than others, causing the model to dedicate most of the 24 hours to such expressions. The percent column in Table 3 show the percent of all corpus occurrences dedicated to each expression. For instance, 81.9% of the occurrences found for Finnish are for night, and the model predicted a 20 hour night. It could be a result of the extremely short days in Finland during the winter, but this is likely exaggerated by the bias in corpus occurrences.

5 Analysis

5.1 Uniformity of Time Distributions

Figure 4 presents the hour distribution for each expression in Italian, as estimated using the extractive (blue) and LM-Dist (orange) methods. As the figure demonstrates, the LM-predicted distribution is more uniform than the extractive one. This is true across most languages: the average entropy of the extractive distributions across languages is 2.78 ± 0.3, and 3.07 ± 0.08 for the LM-Based distributions. For comparison, a uniform distribution across all 24 hours yields an entropy of 3.18.

The uniform distributions predicted by BERT are possibly caused by the similarity between the different inputs (time expressions) and the different outputs (numbers). Previous work showed that BERT confuses semantically-similar but mutually-exclusive concepts such as colors (Shwartz and Choi, 2020). The representation of numbers in distributional models is also suboptimal (Naik et al., 2019; Thawani et al., 2021).

	Morning			Noon			Afternoon			Evening			Night		
	Start	**End**	**%**	**Start**	**End**	**%**	**Start**	**End**	**%**	**Start**	**End**	**%**	**Start**	**End**	**%**
EN	4:00	12:00	36.3	12:00	13:00	6.6	13:00	17:00	11.7	17:00	18:00	16.4	18:00	4:00	29.0
DE	4:00	15:00	34.7	15:00	16:00	6.1	16:00	17:00	8.3	17:00	22:00	20.5	22:00	4:00	30.4
FR	3:00	11:00	35.6	11:00	17:00	21.3	17:00	18:00	1.1	18:00	19:00	10.3	19:00	3:00	31.8
PL	1:00	12:00	55.8	12:00	21:00	29.1	21:00	22:00	2.0	22:00	23:00	1.8	23:00	1:00	11.3
JA	5:00	12:00	41.3	12:00	13:00	6.4	13:00	15:00	6.1	15:00	18:00	8.1	18:00	5:00	38.1
IT	6:00	12:00	24.4	12:00	13:00	4.8	13:00	18:00	20.3	18:00	22:00	20.2	22:00	6:00	30.2
NL	4:00	13:00	31.4	13:00	17:00	17.6	17:00	18:00	2.5	18:00	21:00	24.0	21:00	4:00	24.5
PT	1:00	11:00	31.3	11:00	12:00	4.0	12:00	21:00	39.3	-	-	0.0	21:00	1:00	25.3
ES	3:00	11:00	29.4	11:00	12:00	6.1	12:00	21:00	40.3	-	-	0.0	21:00	3:00	24.2
RU	7:00	11:00	21.6	11:00	13:00	15.4	13:00	14:00	3.4	14:00	15:00	11.5	15:00	7:00	48.0
SV	6:00	11:00	23.7	11:00	12:00	9.4	12:00	13:00	7.5	13:00	22:00	26.8	22:00	6:00	32.6
ZH	6:00	12:00	20.0	12:00	13:00	3.2	13:00	18:00	14.4	18:00	20:00	25.5	20:00	6:00	36.9
NO	7:00	11:00	16.8	11:00	12:00	1.6	12:00	13:00	14.8	13:00	22:00	32.4	22:00	7:00	34.4
FI	12:00	13:00	6.0	13:00	14:00	0.2	14:00	15:00	0.6	15:00	16:00	11.3	16:00	12:00	81.9
CA	4:00	15:00	39.0	15:00	16:00	7.1	16:00	17:00	16.7	17:00	18:00	8.8	18:00	4:00	28.3
UK	8:00	10:00	12.5	10:00	11:00	2.8	11:00	12:00	16.7	12:00	13:00	10.6	13:00	8:00	57.3
TR	4:00	12:00	36.6	12:00	13:00	0.3	13:00	14:00	5.9	14:00	22:00	23.4	22:00	4:00	33.8
CS	1:00	16:00	46.3	16:00	17:00	8.5	17:00	18:00	19.0	18:00	23:00	20.2	23:00	1:00	6.0
HU	3:00	11:00	30.6	11:00	12:00	13.8	12:00	16:00	17.6	16:00	23:00	26.6	23:00	3:00	11.4
ID	4:00	11:00	36.4	11:00	15:00	16.4	15:00	18:00	9.2	-	-	0.0	18:00	4:00	37.9
HE	4:00	11:00	19.7	11:00	12:00	5.6	12:00	18:00	28.6	18:00	22:00	26.2	22:00	4:00	19.9
AR	1:00	2:00	39.7	2:00	3:00	0.2	3:00	4:00	5.7	4:00	23:00	53.5	23:00	1:00	0.9
KO	3:00	4:00	13.1	4:00	5:00	0.8	5:00	10:00	31.9	10:00	11:00	8.3	11:00	3:00	45.9
VI	1:00	12:00	52.9	12:00	13:00	6.6	13:00	18:00	25.8	18:00	19:00	2.3	19:00	1:00	12.5
FA	7:00	11:00	42.0	11:00	12:00	0.0	12:00	20:00	34.6	20:00	21:00	1.2	21:00	7:00	22.2
EL	1:00	11:59	45.1	11:59	15:00	19.9	-	-	0.0	15:00	21:00	23.6	21:00	1:00	11.4
HI	10:00	11:00	35.6	11:00	12:00	0.0	12:00	13:00	16.0	13:00	14:00	0.8	14:00	10:00	47.6

Table 3: Start and end time for various languages, as predicted by the extractive method, along with the percent of corpus occurrences for each expression.

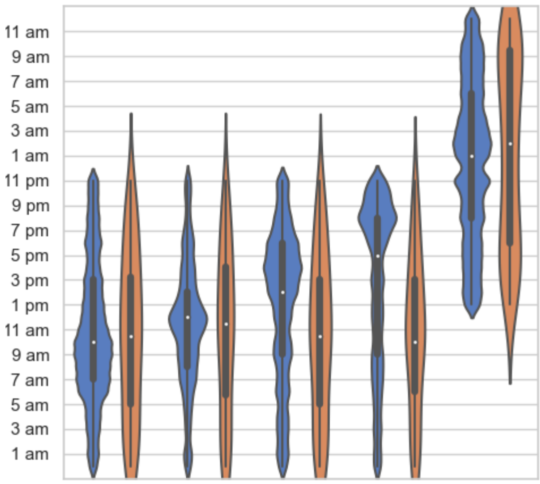

Figure 4: Distribution of hours per time expressions in Italian as estimated by the extractive (blue) and LM-based Dist (orange) methods.

5.2 Analysis of Extracted Sentences

We sample 25 English sentences extracted by the extractive method (§3.1), and examine whether they are valid, manually categorizing the errors. Table 4 presents the percents of each category, along with representative examples. In accordance with the results in Table 2, most of the extractions were valid. Among the errors, 4 sentences contained reference errors, for instance reporting on someone being injured in the morning and dying at another time of the day a few days later. Three sentences included a citation from the Bible or the New Testament, treating the chapter and verse separated by a colon as a time mention.

We repeated the same analysis for languages spoken by members of our research group: Chinese, Korean, Russian, Hebrew, and Italian. The percent of valid sentences ranged from 52% (Chinese) to 80% (Korean). Across languages, reference was a common error in longer paragraphs, but in preliminary experiments we found that splitting the paragraphs to sentences yields a sparse signal. In Chinese, that uses both 12-hour and 24-hour notations, the 12-hour clock was sometimes used without specifying am or pm in unambiguous contexts such as "5:00 in the afternoon". In Hebrew, the word for "evening" has a rarer meaning of "before" which led to WSD error. In Korean, we translated "afternoon" to 오후 that more broadly means "pm".

5.3 Similarity Across Languages

Using the predictions from the extractive method (§3.1), we compute the accuracy of predicting the start and end times of each language from the times of each other language. Figure 5 shows a heatmap of the most similar and most dissimilar languages with respect to time ranges.

The most similar language pairs in terms of time ranges are pairs of closely related languages: Norwegian and Swedish (100%) followed by Por-

Type	%	Example	
① Valid	72%	Every evening at 18:45	
② Reference error	16%	suffered apoplectic fit on the morning of 2 February, and died at 11:45 am, 4 days later	
③ Verse	12%	“Book of Signs” (1:19–12:50); the account of Jesus’ final night	
④ 12-hr clock without am/pm		下午1:00-5:00開放	Between 1:00-5:00 in the afternoon.
⑤ WSD error		ערב המלחמה... בשעה 17:00 הגיע הכוח	Before the war... at 17:00, the force arrived
⑥ Imperfect time expression mapping		매주 토요일, 오후 19:00-21:30.	Every Saturday at 19:00-21:30 pm.

Table 4: **Top**: Manual categorization of a sample of the English sentences extracted in the extractive method, along with a (slightly shortened) example of each category. **Bottom**: additional error examples in other languages.

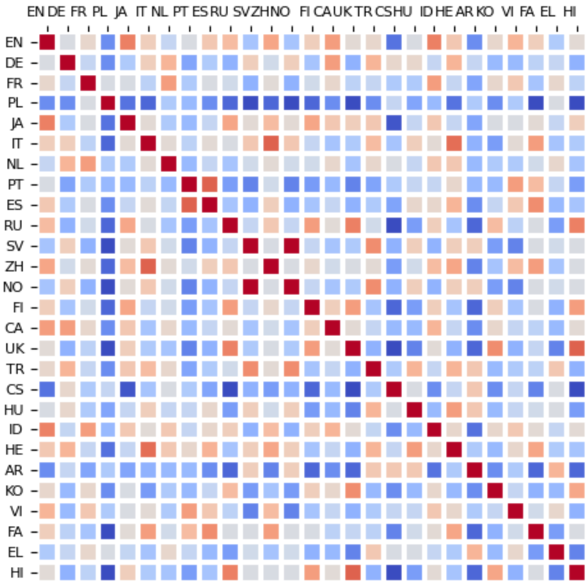

Figure 5: Heatmap indicating the accuracy of predicting the start and end times of each language from the times of each other language. Dark red indicates 100% accuracy while dark blue indicates 0% accuracy.

tuguese and Spanish (92%). In particular, the latter two don’t distinguish evening from night.

The similarity between Italian and Chinese (92%) might be explained by the similarity between the average times of waking up and going to bed in both countries: both Italian men and Chinese women go to sleep close to midnight and wake up around 7:30 on average (Walch et al., 2016).

Finally, Hindi and Ukrainian have similar predictions as well (92%), but considering the extremely early night start time predicted for both (2 pm and 1 pm), we conjecture that this is mostly due to noise in the data. The same pattern emerges between pairs of dissimilar languages such as Czech and Russian or Farsi and Polish (36%), where the model of each language devotes most of its 24 hours to a single time expression.

6 Related Work

Temporal Commonsense. Work on temporal reasoning ranges from extracting and normalizing temporal expressions (Strötgen and Gertz, 2010; Angeli et al., 2012; Vashishtha et al., 2019), to inferring possibly explicit temporal attributes of events, including their order (Ning et al., 2018; Vashishtha et al., 2019), duration (Chambers and Jurafsky, 2008; Vashishtha et al., 2019), and typical times or frequencies (Zhou et al., 2019).

Various benchmarks were proposed to measure models’ temporal reasoning abilities. The bAbI suite contains a task that requires reasoning about the order of time expressions (Weston et al., 2015). MC-TACO is a reading comprehension task pertaining to ordering, duration, stationarity, frequency, and typical times of events (Zhou et al., 2019). TIMEDIAL (Qin et al., 2021) is a dialogue QA task focusing on temporal commonsense. Zhou et al. (2021) and Thukral et al. (2021) both cast the temporal ordering task as an NLI task. In another line of work, tracking state changes in procedural text is also related to temporal ordering (Dalvi et al., 2018; Zhang et al., 2020). Despite the success of pre-trained LMs on language understanding tasks, their performance on these benchmarks is limited, maybe due to the fact that many temporal relations are not explicitly stated in text (Davis and Marcus, 2015). A promising direction is to train LMs explicitly on temporal knowledge (Zhou et al., 2020).

Cultural Commonsense. There is little focus on cultural differences in NLP in general (Hovy and Yang, 2021) and in research about commonsense reasoning in particular. Recently, Acharya et al. (2021) made a first step in addressing this gap. They surveyed crowdsourcing workers in the US and India regarding rituals that are commonly found across cultures such as birth, marriage, and funerals. In particular, they asked questions pertaining to temporal aspects such as typical time and duration of each event. The paper mentions anecdotal differences such that a wedding lasts a few hours in the US but a few days in India.

Although there is no direct mapping between culture and language, one can often teach about the other. For example, in ConceptNet, a multilingual commonsense knowledge base, the English entry for breakfast specifies pancakes as breakfast food, while the Chinese entry mentions noodles (Speer et al., 2017).

Language Grounding and World Knowledge. Our work is related to language grounding (Roy and Reiter, 2005) and to extracting world knowledge from text corpora (Carlson et al., 2010; Tandon et al., 2014). In the intersection of these two lines of work, Forbes and Choi (2017) extracted from a corpus physical commonsense knowledge about actions and objects along five dimensions (size, weight, strength, rigidness, and speed), while Elazar et al. (2019) induced distributions of typical values of various quantitative attributes such as time, duration, length, and speed. Elazar et al. (2019) mention cultural differences that arose when crowdsourcing workers were asked to estimate whether an item's price was expensive or not: annotators from India judged prices differently from annotators in the US.

7 Discussion and Conclusion

We addressed the task of grounding time expressions such as "morning" and "noon" in different languages to explicit hours. Our extractive method achieves good performance on languages for which we collected gold annotations. We dedicate the remainder of the paper to discuss various limitations and considerations for future work.

Temporal and Seasonal Factors. As discussed in §2.3, some workers mentioned that their interpretation of time expressions depends on the season, e.g., night starts earlier in the winter in the Northern Hemisphere. In addition, the time of day in which the workers answered the survey might have introduced some bias. The batches were published according to the authors' timezone and working hours, which might have been outside working hours for some countries. An early riser answering an AMT survey at 5 am or a night owl that answers it at 2 am might not be representative of the population. Finally, Vilares and Gómez-Rodríguez (2018) showed that tweets greeting "good morning" appeared later in the day during weekends and holidays, indicating later wake up times. It is possible that such factors will also affect the judgement of survey respondents.

Reporting Bias. Every method that learns about the world from texts (or from language models, trained on text corpora), suffers from reporting bias (Gordon and Van Durme, 2013; Shwartz and Choi, 2020). The frequency of occurrences in a corpus is an imperfect proxy for measuring the quantity or frequency of things in the world. In our case, it may be that some hours are less spoken of in general: perhaps fewer newsworthy events happen late at night? Some time expressions might be less ambiguous than others and therefor appear less frequently with an exact time mention.

Inducing time distributions from greetings also confounds other cultural factors such as politeness. The mapping between greetings and time expressions is not perfect, e.g. as Vilares and Gómez-Rodríguez (2018) note, "bonjour" in French means "good morning" but is also used throughout the day to mean "hello". Finally, Twitter memes might use a greeting with a different intention, as in the famous "good morning to everyone except" meme.[7]

Differences in Performance across Languages. While the methods in this paper are language-agnostic, they don't produce equally good predictions for all languages. Beyond the differences in the set of commonly used time expressions in each language (e.g., "evening" being missing from Spanish, or "dawn" being commonly used in other languages), time might also be discussed differently in different languages. In some languages it may be more common to use cardinals to discuss hours, as in "It is two in the afternoon". Finally, the success of our methods also depends on the availability of large text corpora and the quality of the LM. We used mBERT because it is available for 104 languages, but we focused on relatively high-resource languages. This model doesn't perform equally well across all languages (Wu and Dredze, 2020). In the future, we plan to find alternative sources for collecting gold standard annotations for additional languages, which will facilitate evaluating the performance of our methods on a broader range of languages.

[7] For instance, several tweets from early 2021 with the hashtag #FreeBritney read "Good morning to everyone except Jamie Spears."

References

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2021. Towards an atlas of cultural commonsense for machine reasoning. In *AAAI*.

Gabor Angeli, Christopher Manning, and Daniel Jurafsky. 2012. Parsing time: Learning to interpret time expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 446–455, Montréal, Canada. Association for Computational Linguistics.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI conference on artificial intelligence*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

AL López. 2015. Infographic: A world of languages-and how many speak them. retrieved november 8, 2015.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380, Florence, Italy. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

7066–7076, Online. Association for Computational Linguistics.

Ehud Reiter and Somayajulu Sripada. 2002. Squibs and discussions: Human variation and lexical choice. *Computational Linguistics*, 28(4):545–553.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1):1–12. Connecting Language to the World.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2016. Identification, characterization, and grounding of gradable terms in clinical text. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 17–26, Berlin, Germany. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2014. Acquiring comparative commonsense knowledge from the web. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing language models for understanding of temporal expressions. In *Blackbox NLP workshop*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2018. Grounding the semantics of part-of-day nouns worldwide using Twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 123–128, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Olivia J Walch, Amy Cochran, and Daniel B Forger. 2016. A global quantification of "normal" sleep schedules using smartphone data. *Science advances*, 2(5):e1501705.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.