

Dataset and Baseline System for Multi-lingual Extraction and Normalization of Temporal and Numerical Expressions

Sanxing Chen*
Duke University
sanxing.chen@duke.edu

Yongqiang Chen*
The Chinese University
of Hong Kong
yqchen@cse.cuhk.edu.hk

Börje F. Karlsson
Microsoft Research Asia
borjekar@microsoft.com

Abstract

Temporal and numerical expression understanding is of great importance in many downstream Natural Language Processing (NLP) and Information Retrieval (IR) tasks. However, much previous work covers only a few sub-types and focuses only on entity extraction, which severely limits the usability of identified mentions. In order for such entities to be useful in downstream scenarios, coverage and granularity of sub-types are important; and, even more so, providing resolution into concrete values that can be manipulated. Furthermore, most previous work addresses only a handful of languages. Here we describe a multi-lingual evaluation dataset - NTX - covering diverse temporal and numerical expressions across 14 languages and covering extraction, normalization, and resolution. Along with the dataset we provide a robust rule-based system as a strong baseline for comparisons against other models to be evaluated in this dataset. Data and code will be publicly available at <https://aka.ms/NTX>.

1 Introduction

Entity recognition (or entity extraction) is a key component in many NLP pipelines and important for various downstream tasks. However, most ER works focus only on *named* entities with types like Person, Organization, etc.; treating other potentially important terms (like datetime mentions or numerals) as only literals.

This is problematic as such entities play important roles in information retrieval, relationship extraction, conversational language understanding, task completion, knowledge base construction, and beyond (Alonso et al., 2007; Grudin and Jacques, 2019; Gesese et al., 2021).

Furthermore, even when covering numerical and temporal entities, datasets treat them in a

too unbalanced or coarsely way - e.g., OntoNotes 5 (Weischedel et al.) has five categories for numerical entities, but only two for dates and times. On the numerical types side, such categorization ignores that numeric literals often denote specific units of measurements or types in context. E.g., "She is eight", clearly implies that "eight" is not just a cardinal number, but a description of *age*.

Meanwhile, while there has been an increasing interest in temporal entities (UzZaman et al., 2013), most efforts utilize complex annotations tagging both datetime expressions per se and also temporal relationships. Such coupled complex annotation schemas both are too intricate to be used effectively by downstream users and do not cover many mention forms necessary in practice.

Due to the complexity of existing annotation schemas and the cost of data acquisition and annotation, quality assets even for the entity types covered only exist in English and sparsely in a handful of languages.

Another parallel issue affecting numerical and temporal entities or expressions is that for them to be of actual use for downstream tasks, mention detection is not enough. Such mentions need to be normalized (turned into a canonical form that standardizes interpretation) and further resolved into concrete values for usage (sometimes necessitating extra context for resolution). E.g. "from February to the end of 2012" -> (XXXX-02,2012-EOY,P11M) -> start: 2012-02-01, end: 2012-12-31. The latter allows the conversion into specific datetime object instances that can be consumed in conventional programs.

Some of the previous efforts in creating schemas for temporal mentions, e.g., UzZaman et al. (2013), also cover normalization/resolution. But they suffer from the same complexity and coverage problems mentioned, or also from mixing normalization and resolution. We emphasize the requirement to separate the two as resolution will in many

*The work described in this technical report was performed during the authors' internships at Microsoft Research Asia.

cases require extra context and identifying this context is error prone. So any system addressing temporal entities should allow re-resolution from the normalized form. For example, "now" would have normalized form "PRESENT_REF". But to turn it into a concrete resolution, an anchor reference datetime such as "2022-07-01" is needed. As such references can be incorrectly inferred, systems should allow re-resolution. A normalized form like "PRESENT_REF" allows it, while the typical annotation in TIMEML would mix the two and directly consider the time expression as "2022-07-01" and re-resolution is not possible without re-processing the original text.

Due to the mentioned limitation of previous efforts and lack of existing datasets, here we propose a multi-lingual benchmark dataset that covers numerical and temporal entities and expressions at the extraction, normalization, and resolution levels.

This dataset has been manually annotated in the context of a multi-year effort over real-world use in commercial applications¹. The proposed schema focuses on real-world entity coverage and ease of use for downstream applications and includes: 8 numerical sub-types (cardinal, ordinal, percentage, numerical range, age, currency, dimension, temperature) and 10 temporal sub-types (date, time, datetime, date range, time range, datetime range, holiday, duration, timezone, and recurring set).

The NTX dataset covers 14 languages², which belong to a diverse set of language families—Indo-European, Sino-Tibetan, Japonic, Turkic, Semitic, and Koreanic.

Furthermore, to both serve as a baseline for performance comparisons over this dataset and to help generate training data for new models following its schema, we reference a high quality rule-based system (Huang et al., 2017) that achieves strong results and has been hardened through real-world commercial use.

2 Related Work

2.1 Numerical Entities

Already in the nineties venues like MUC attempted to standardize the evaluation of information extraction (IE) tasks, including numerical expression extraction (and in six languages) (Palmer and Day, 1997). Perhaps due to high coverage of simple

¹The Recognizers-Text project, available at <https://github.com/microsoft/Recognizers-Text>.

²See detailed list of languages in Section 3.

rules for dataset cases at the time, most typical entity recognition datasets do not include numerical types (e.g., CoNLL 2002/2003).

Later datasets like OntoNotes (Weischedel et al.), recognize the importance of such entities (e.g., Percent, Money, Quantity, Ordinal, and Cardinal). But such inclusion is not commonplace and annotation is still restricted only to tagging/extraction.

As more complex tasks from QA to document understand get traction, interest has shined again on numerical entities and their importance, especially in domains such as finance and healthcare. This is evidenced by both new tasks like the FinNum series (Chen et al., 2019a, 2020, 2021) and works in Health IE (Jagannatha et al., 2019).

However, such newer datasets are too domain-specific (e.g., buy price, sell price, and stop loss in FinNum) or target only extraction. Not to mention a lack of consistency between annotation types.

Moreover semantics encoded in numerical entities both capture type information and often denote units of measurements. Considering only the numeric value results in loss of knowledge (e.g., magnitude or sub-type compatibility) (Gesese et al., 2021). Other challenges also include variability in mention forms, un-anchored ordinals, and range expressions. Ideally all of which should be interpretable as they significantly change the semantics of a given mention (e.g., "one"; "30+"; "half").

Alternate works like AMR (Banarescu et al., 2013) also recognize the importance of representing and tagging numerical and temporal terms (e.g., :quant, :unit, :year, :season, :weekday, etc.), but do not address typing and purposefully do not perform any normalization. Moreover, their 50-pages annotation guidelines are overly complex.

2.2 Temporal Entities

Differently from numerical expressions, there have been frequent efforts in creating corpora and assessing their quality regarding temporal entities. Such as the TempEval series (Verhagen et al., 2007, 2010; UzZaman et al., 2013) and annotated corpora like ACE³ and TimeBank⁴.

However, even with a certain degree of maturity, most annotation approaches suffer from low coverage of mention forms, too high complexity, difficulty to extend, and lack of granularity for downstream use. The most popular annotation

³LDC2005T07 and LDC2006T06 in the LDC catalogue.

⁴LDC2006T08 in the LDC catalogue.

standards are TIDES TIMEX2 (Ferro et al., 2005) and TimeML’s TIMEX3 (Pustejovsky et al., 2005), which are used in datasets like TempEval and WikiWars (Mazur and Dale, 2010).

The limitations of TIMEX2, for example, span the annotation of time zones, event-based expressions, duration and set anchor restrictions.⁵ Moreover annotation guidelines for such schemas are complex, abstract, and sometimes open to ambiguous interpretation (Saurí et al., 2006).

Previous attempts to address temporal expression recognition and resolution have also highlighted other limitations of such schemas. From problems in the standard evaluation datasets (Li et al., 2014) (with missing and incorrect annotations), to specific cumbersome annotation requirements such as "Empty tags are TIMEX3 tags that do not contain any tokens and should be created whenever a temporal expression can be inferred from preexisting text-consuming TIMEX3 tags" which is either not applied or inconsistently done (Manfredi et al., 2014), leading to issues with anchored durations (e.g., "a month ago") and range expressions that combine two TIMEX3 tags (e.g., "from 2010 to 2014").

Support for temporal ranges in general is non-intuitive. For instance, expressions like "from 3 to 4 p.m." to "12-13 March 2011" are hard or impossible to annotate for their ambiguities, and when annotated or generated in the response of extraction, are hard to be used downstream.

In order to reduce complexity and cover more mention forms in an easy to consume way, our time expressions differ from TIMEX2/TIMEX3. Mostly in the granularity of types and representation of complex expressions like durations and recurring datetimes.⁶

Variety of mention forms highlight also the need for consistency between numerical and temporal expression recognition. Time will depend on numbers. For example "in half an hour" requires consistent handling of fraction term and articles.

Lastly, many common mention forms like "8:24 a.m. Chicago time" are not well covered by previous guidelines, but are covered in NTX.

⁵Sentences like "every Tuesday since March" or "five days in mid-August" can not be precisely annotated

⁶E.g., TimeML tags "November" as Date, while NTX tags it as DateRange

3 Dataset Details

To alleviate some of the described issues and trying to cover a wide variety of scenarios we propose a new dataset - NTX (Numerical and Temporal eXpressions) - for the evaluation of numerical and temporal recognition systems.

NTX was build on real usage over the past several years and targets cross-domain scenarios and the interrelated nature of numex and timex. Coverage includes variants of languages (e.g., French covers both fr-FR and fr-CA) and formal and informal mention forms.

The dataset covers 14 languages - English, Chinese, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Spanish, Swedish, Turkish, Hindi, and Arabic; which belong to a diverse set of language families. With 8 sub-types of numerical entities (cardinal, ordinal, percentage, numerical range, age, currency, dimension, and temperature) and 10 temporal sub-types (date, time, datetime, date range, time range, datetime range, holiday, duration, timezone, and recurring set).

These entities both provide fine-grained granularity, many times required by downstream tasks, and help address the previously mentioned limitations of other datasets.

This is accomplished mainly in two fronts: i) Allowing fine-grained types that keep semantics useful in downstream tasks (i.e., not mixing the concepts of date, time, and ranges) and adding new sub-types for previously not supported annotations (e.g., such as holidays for mentions like "Xmas", "Easter Sunday", etc.); and ii) Simplifying annotation by, instead of having complex multi-level annotations (entities, relationships, and modifiers annotated in different ways), grouping them as much as possible and representing them in a streamlined entity-level form (i.e., permit combination of modifiers, representing a range by start/end/length, instead of complex relationships). For example, "after mid-August" is 1 entity (of date range type), and not 3 entities plus additional relationship annotations.

The dataset contains over 26000 sentences across the different languages.

To fulfil its requirement of covering multiple scenarios, it includes both long and short sentences. Specifically to cover common conversational scenarios, where lack of context is commonplace, approximately 13% of cases consist of only an entity mention (i.e., they could be the input directly to normalization and resolution as-is).

	AR	ZH	NL	EN	FR	DE	HI	IT	JA	KO	PT	ES	SV	TR
w/ numerical entities	526	1068	433	1127	696	207	582	436	1442	713	647	691	394	457
w/ temporal entities	909	518	2064	1920	2500	480	1368	808	1496	1004	526	1660	461	962
Overall	1762	2007	2778	3546	3677	769	2229	1434	3796	2057	1409	2661	1043	1564

Table 1: Numbers of sentences with entities by language. *Overall* includes sentences with no entity.

```
{
  "Input": "The number is between 20 and 30.",
  "Results": [{
    "Text": "between 20 and 30",
    "TypeName": "numberrange",
    "Resolution": {
      "value": "[20,30]"
    },
    "Start": 14,
    "End": 30
  }]
}
```

Figure 1: Number range annotation.

```
{
  "Input": "Help me book a meeting tonight from 7 to 7:30pm",
  "Context": {
    "ReferenceDateTime": "2016-11-07T10:20:00",
    "Culture": "en-US"
  },
  "Results": [{
    "Text": "tonight from 7 to 7:30pm",
    "Start": 23,
    "End": 46,
    "TypeName": "datetimev2.datetimerange",
    "Resolution": {
      "values": [{
        "timex": "(2016-11-07T19,2016-11-07T19:30,PT30M)",
        "type": "datetimerange",
        "start": "2016-11-07 19:00:00",
        "end": "2016-11-07 19:30:00"
      }]
    }
  }]
}
```

Figure 2: DateTime range annotation.

Moreover, the dataset includes not only sentences with entity mentions, but also multiple sentences that correctly have no annotation. This is necessary to make sure evaluations also cover behaviour related to false positives. Table 1 shows a summary of sentence counts in the dataset⁷.

Data is made available in the form of JSON files in order to represent not only the tagging of entities, but also a representation of mention normalization and potential resolutions. Figures 1 and 2 show two example mentions.

Nonetheless, resolution of relative expressions or un-anchored mentions can still be ambiguous. In such cases the dataset lists both future and past resolutions as acceptable. For example, "Friday"

can be interpreted either as "upcoming Friday" or "past Friday".

Details of the dataset creation are provided in Appendix A.

4 Rule-Based System Design

While the limitations of rule-based systems are well known, especially in regards to the maintenance of large amounts of rules and their interactions, we have opted in Recognizers-Text (Huang et al., 2017) for a rule-based design due to three key design principles:

i) determinism: the system needs to always produce predictable output, so downstream consumers of it's output can act accordingly; ii) prioritize recall: as rules are prone to false positives or false negatives, the system should focus on coverage to the extent possible, as false positives could potentially be filtered in pre- or post-processing stages. iii) no need for expert knowledge to make changes: rules are a somewhat straightforward way to represent knowledge for entity extraction, instead of requiring users to have a linguistic or machine learning background.

Although neural architectures have shown to be able to perform competitively on time expression recognition on previous datasets (Lange et al., 2020; Chen et al., 2019b; Cao et al., 2022), such architectures don't address the requirements above, and both their implementation and evaluation target only recognition.

The core structure of the rule-based extractors is inspired by SUTIME (Chang and Manning, 2012) and basically follows its *three-types-of-rules* design. Rules are roughly categorized into mention capture, composition, and filtering. The same structured core of rules is shared across languages while localized for language-specific properties. We utilize regular expressions throughout the system.

For better performance, to avoid unwieldy long regexes, the currency and timezone extractors also make use of dictionaries, in the form of tries (prefix trees) for tagging. This has the additional benefit of facilitating users scenarios where they require to load their own extra terms.

⁷Detailed statistics by sub-type are shown in Tables 2 and 3 in Appendix B.

Normalization and resolution (together termed parsing in our system) however, are key, and require additional code writing. Parsing takes the form of a cascade of parsers per type increasing in complexity.

Language-specific behaviour is defined as overridable functions in each language configuration. If a function is not overridden, the core default behaviour is adopted (with other language-specific configurations).

5 Conclusion

Here we propose NTX, a novel multi-lingual evaluation dataset covering diverse temporal and numerical expressions across 14 languages. We also provide Recognizers-Text as a robust baseline system for comparisons against other models over this evaluation dataset.

Limitations

As NTX includes not only span detection, but also type information and resolution, potential ambiguities in type are important. Where ambiguities were detected during the dataset creation, consensus of expert annotators was used to determine the most likely (most commonly applicable) type. We plan for future versions of the dataset to include information about such alternative type interpretations.

The described rule-based system (Recognizers-Text) is intended primarily as a strong evaluation baseline over NTX for performance comparisons, but it can also serve as a potential source of automatically labeled data in the NTX schema for training semi-supervised models⁸. As mentioned in Section 4, rule-based systems have well known limitations, including handling of false positives. Utilizing the described baseline system to automatically generate annotated data must be followed by annotation review to account for such issues.

It is also important to note that, while the dataset contains the data for temporal expressions in SV, KO, and AR, as well as numerical expressions with units in AR, these are currently not supported by the baseline rule-based system.

Lastly, the current dataset schema may be somewhat unintuitive to manual inspection; as it focuses on extraction/parsing representation. The open-

sourced code includes evaluation scripts to calculate Precision/Recall/F-1 over it, for ease of use.

References

- Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2007. [On the value of temporal information in information retrieval](#). *SIGIR Forum*, 41(2):35–41.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Yuwei Cao, William Groves, Tanay Kumar Saha, Joel Tetreault, Alejandro Jaimes, Hao Peng, and Philip Yu. 2022. [XLTime: A cross-lingual knowledge transfer framework for temporal expression extraction](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1931–1942, Seattle, United States. Association for Computational Linguistics.
- Angel X. Chang and Christopher Manning. 2012. [SU-Time: A library for recognizing and normalizing time expressions](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).
- Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019a. Final report of the NTCIR-14 finnum task: Challenges and current status of fine-grained numeral understanding in financial social media data. In *NII Testbeds and Community for Information Access Research - 14th International Conference, NTCIR 2019, Tokyo, Japan, June 10-13, 2019, Revised Selected Papers*, volume 11966 of *Lecture Notes in Computer Science*, pages 183–192. Springer.
- Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the ntcir-15 finnum-2 task: Numeral attachment in financial tweets. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, pages 75–78.
- Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2021. Overview of the ntcir-16 finnum-3 task: Investor’s and manager’s fine-grained claim detection. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*.
- Sanxing Chen, Guoxin Wang, and Börje F. Karlsson. 2019b. [Exploring word representations on time expression recognition](#). Technical Report MSR-TR-2019-46, Microsoft Research.

⁸Different implementations of the system are available in the GitHub repository, which may present differing output quality. The .NET version is recommended as the canonical version for evaluations.

- Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. Tides: 2005 standard for the annotation of temporal expressions. Technical report, MITRE CORP MCLEAN VA.
- Genet Asefa Gesese, Russa Biswas, Mehwish Alam, and Harald Sack. 2021. A survey on knowledge graph embeddings with literals: Which model links better literal-ly? *Semantic Web*, 12(4):617–647.
- Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA. Association for Computing Machinery.
- Wenhao Huang, Zijia Lin, Chris McConnell, and Börje F. Karlsson. 2017. Recognizers-Text: Recognition and resolution of numbers, units, and date/time entities expressed across multiple languages.
- Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.
- Lukas Lange, Anastasiia Iurshina, Heike Adel, and Janik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 103–109, Online. Association for Computational Linguistics.
- Hui Li, Janik Strötgen, Julian Zell, and Michael Gertz. 2014. Chinese temporal tagging with heideltime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 133–137.
- Giulio Manfredi, Janik Strötgen, Julian Zell, and Michael Gertz. 2014. Heideltime at eventi: Tuning italian resources and addressing timeml’s empty tags. *HeidelTime at EVENTI: Tuning Italian Resources and Addressing TimeML’s Empty Tags*, pages 39–43.
- Pawet Mazur and Robert Dale. 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, page 913–922, USA. Association for Computational Linguistics.
- David D. Palmer and David S. Day. 1997. A statistical profile of the named entity task. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, page 190–193, USA. Association for Computational Linguistics.
- James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2/3):123–164.
- Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. Timeml annotation guidelines, version 1.2.1.
- Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.
- Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. [OntoNotes Release 5.0. LDC2013T19.](#)

		Language													
		Arabic	Chinese	Dutch	English	French	German	Hindi	Italian	Japanese	Korean	Portuguese	Spanish	Swedish	Turkish
Sub-type	Cardinal	284	362	162	233	273	55	192	152	764	254	174	206	152	129
	Number Range	145	65	53	87	85	6	76	34	185	132	50	98	30	61
	Ordinal	70	7	48	54	70	38	64	31	101	68	33	82	43	39
	Percentage	27	152	13	20	35	15	16	11	204	30	51	66	11	11
	Age	0	10	17	19	18	14	21	15	18	19	18	18	20	18
	Currency	0	35	32	180	114	39	115	104	68	108	124	122	36	109
	Dimension	0	36	74	93	60	28	53	55	64	64	65	54	67	56
	Temperature	0	10	34	36	41	12	45	34	38	38	47	45	35	34

Table 2: Numbers of sentences with numerical entities by language and sub-type.

		Language													
		Arabic	Chinese	Dutch	English	French	German	Hindi	Italian	Japanese	Korean	Portuguese	Spanish	Swedish	Turkish
Sub-type	Date	118	64	226	148	228	28	122	107	214	194	65	73	132	94
	Date Range	305	67	319	352	514	63	318	241	246	180	60	374	60	221
	DateTime	72	15	134	81	125	21	67	68	87	59	61	63	15	64
	DateTime Range	73	27	180	96	183	32	86	77	99	62	41	118	27	67
	Model - Overall	112	257	741	916	884	202	507	114	584	316	156	825	112	316
	Duration	55	15	87	62	103	21	61	44	44	34	21	23	15	33
	Holiday	18	35	46	26	50	61	30	11	43	15	14	17	11	19
	Set	27	8	52	32	58	13	30	27	33	27	20	18	8	25
	Time	74	14	138	93	198	26	84	67	79	61	58	56	14	69
	Time Range	55	16	122	65	109	13	63	52	67	56	30	93	13	54
	Timezone	0	0	19	49	48	0	0	0	0	0	0	0	54	0

Table 3: Numbers of sentences with temporal entities by language and sub-type.

A Dataset creation process

The dataset was created firstly in English and Chinese by using hired specialist vendors with linguistic expertise to generate sentences and utterances covering formal and informal cases in both document-type sentences, as well as, conversational/social media-type utterances. The same vendors (level 1 annotators) were tasked with annotating all entity types in the collected data. Upon completion of this stage, two expert annotators (level 2 annotators) validated all annotations and documented existing disagreements and desired behaviour on possibly ambiguous cases.

The baseline rule-based system was developed in parallel and was used to further validate annotation. Any failure cases of the system execution resulted in either corrections in the annotations or in extensions to the system to properly address them. Cases where the annotations are agreed to be correct, but that cannot yet be supported by the system are marked as such in metadata to not break a system run during development (but still included in the dataset).

Language expansion then happened through a mixed process of i) generating data in a similar fashion to English and Chinese, and ii) translating large numbers of examples from the existing languages into new target languages (creating parallel texts). Translation was performed by native speaker vendors. Such process had the benefit of emphasizing a balance between language specific mention formats in generation, while having a certain common coverage across languages via translation.

Moreover, the datasets in each language grew organically through long time usage of the system in a commercial setting, collecting failure feedback cases along with support requests for new scenarios (both added as new examples on the dataset).

B Dataset statistics

Here we provide detailed sets of statistics on the current dataset version.

The distribution of sub-types per language is shown in Tables 2 and 3. Table 4 shows statistics on sentence length and amount of annotated entities per sentence across languages.

		Arabic	Chinese	Dutch	English	French	German	Hindi
Sentences	Distinct	1380	1141	3014	3133	4410	809	2228
	Avg. Length	25.83	14.24	38.04	39.40	42.66	38.31	39.89
	Stdev	14.22	10.00	25.71	30.05	35.24	33.61	31.11
Entities	Total	1866	1577	3614	5254	5693	949	2626
	Average	1.03	1.07	1.06	1.11	1.09	1.06	1.05
	Stdev	0.19	0.32	0.29	0.40	0.36	0.34	0.25
		Italian	Japanese	Korean	Portuguese	Spanish	Swedish	Turkish
Sentences	Distinct	1424	2546	1747	1240	2509	425	1678
	Avg. Length	40.89	17.10	23.89	41.67	43.00	42.40	37.61
	Stdev	39.79	12.50	17.51	41.58	35.31	39.18	33.86
Entities	Total	1837	3275	2112	1386	2903	549	2159
	Average	1.05	1.07	1.05	1.04	1.06	1.07	1.05
	Stdev	0.25	0.30	0.24	0.28	0.30	0.29	0.26

Table 4: Statistics of sentence length and entity appearances per language.

		Language													
		Arabic	Chinese	Dutch	English	French	German	Hindi	Italian	Japanese	Korean	Portuguese	Spanish	Swedish	Turkish
Entities	Timex Counts	1618	843	3307	4279	5166	780	2197	1539	2502	1639	854	2393	326	1799
	Numex Counts	248	734	307	975	527	169	429	298	773	473	532	510	223	360
	Total Counts	1866	1577	3614	5254	5693	949	2626	1837	3275	2112	1386	2903	549	2159

Table 5: Distribution of numerical and temporal entities.

	# Sentences	# Timex Entities	# Numex Entities
TempEval-3 (train)	3987	1822	0
TempEval-3 (test)	273	138	0
Tweets (train)	1662	892	0
Tweets (test)	422	237	0
Wikiwars (train)	3822	2278	0
Wikiwars (test)	1537	373	0
OntoNotes 5.0 (train)	59924	12155	13861
OntoNotes 5.0 (dev)	8528	1721	1721
OntoNotes 5.0 (test)	8262	1814	1898
NTX (English)	3133	4279	975

Table 6: Overall statistics of different datasets (English).

	Avg. Length	# Timex Entities				# Numex Entities				
		Date	Set	Duration	Time	Cardinal	Money	Ordinal	Percent	Quantity
TempEval-3 (train)	21.55	1505	30	257	30	-	-	-	-	-
TempEval-3 (test)	22.61	96	4	34	4	-	-	-	-	-
Tweets (train)	7.38	554	32	167	139	-	-	-	-	-
Tweets (test)	8.01	164	6	33	34	-	-	-	-	-
Wikiwars (train)	22.51	1992	19	200	67	-	-	-	-	-
Wikiwars (test)	21.66	330	4	22	17	-	-	-	-	-
OntoNotes 5.0 (train)	18.16	10922	-	-	1233	7367	2434	1640	1763	657
OntoNotes 5.0 (dev)	17.32	1507	-	-	214	938	274	232	177	100
OntoNotes 5.0 (test)	18.49	1602	-	-	212	935	314	195	349	105

Table 7: Detailed statistics of different datasets (English).

Moreover, Table 5 shows the overall distribution of numerical and temporal entities in the dataset. While Tables 6 and 7 show a comparison of NTX to other common datasets with numeric and temporal expressions and their respective testing splits.