Classification of telicity using cross-linguistic annotation projection

Annemarie Friedrich[1] **Damyana Gateva**[2]
[1]Center for Information and Language Processing, LMU Munich
[2]Department of Computational Linguistics, Saarland University
anne@cis.uni-muenchen.de dgateva@coli.uni-saarland.de

Abstract

This paper addresses the automatic recognition of telicity, an aspectual notion. A *telic* event includes a natural endpoint (*she walked home*), while an *atelic* event does not (*she walked around*). Recognizing this difference is a prerequisite for temporal natural language understanding. In English, this classification task is difficult, as telicity is a covert linguistic category. In contrast, in Slavic languages, aspect is part of a verb's meaning and even available in machine-readable dictionaries. Our contributions are as follows. We successfully leverage additional silver standard training data in the form of projected annotations from parallel English-Czech data as well as context information, improving automatic telicity classification for English significantly compared to previous work. We also create a new data set of English texts manually annotated with telicity.

1 Introduction

This paper addresses the computational modeling of *telicity*, a linguistic feature which represents whether the event type evoked by a sentence's verb constellation (i.e., the verb and its arguments and modifiers) has a natural endpoint or not (Comrie, 1976; Smith, 1997), see (1a) and (1b).

(1) (a) Mary ate an apple. (*telic*)
(b) I gazed at the sunset. (*atelic*)

Automatic recognition of telicity is a necessary step for natural language understanding tasks that require reasoning about time, e.g., natural language generation, summarization, question answering, information extraction or machine translation (Moens and Steedman, 1988; Siegel and McKeown, 2000). For example, there is an entailment relation between English Progressive and Perfect constructions (as shown in (2)), but only for atelic verb constellations.

(2) (a) He was swimming in the lake. (*atelic*)
$\models$ He has swum in the lake.
(b) He was swimming across the lake. (*telic*)
$\not\models$ He has swum across the lake.

We model telicity at the word-sense level, corresponding to the *fundamental aspectual class* of Siegel and McKeown (2000), i.e., we take into account the verb and its arguments and modifiers, but no additional aspectual markers (such as the Progressive). In (2) we classify whether the event types "swim in the lake" and "swim across the lake" have natural endpoints. This is defined on a linguistic level rather than by world knowledge requiring inference. "Swimming in the lake" has no natural endpoint, as also shown by the linguistic test presented in (2). In contrast, "swimming across the lake" will necessarily be finished once the other side is reached.

In English, the aspectual notion of telicity is a covert category, i.e., a semantic distinction that is not expressed overtly by lexical or morphological means. As illustrated by (2) and (3), the same *verb type* (lemma) can introduce telic and atelic events to the discourse depending on the context in which it occurs.

(3) (a) John drank coffee. (*atelic*)
(b) John drank a cup of coffee. (*telic*)

In Slavic languages, aspect is a component of verb meaning. Most verb types are either *perfective* or *imperfective* (and are marked as such in dictionaries). For example, the two occurrences of "drink" in (3) are translated into Czech using the imperfective verb "pil" and the perfective verb

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2542–2548
Copenhagen, Denmark, September 7–11, 2017.

"vypil," respectively (Filip, 1994):[1]

(4) (a) Pil kávu. (*imperfective*)
He was drinking (some) coffee.
(b) Vypil kávu. (*perfective*)
He drank up (all) the coffee.

Our contributions are as follows: (1) using the English-Czech part of InterCorp (Čermák and Rosen, 2012) and a valency lexicon for Czech verbs (Žabokrtský and Lopatková, 2007), we create a large silver standard with automatically derived annotations and validate our approach by comparing the labels given by humans versus the projected labels; (2) we provide a freely available data set of English texts taken from MASC (Ide et al., 2010) manually annotated for telicity; (3) we show that using contextual features and the silver standard as additional training data improves computational modeling of telicity for English in terms of F_1 compared to previous work.

2 Related work

Siegel and McKeown (2000, henceforth SMK2000) present the first machine-learning based approach to identifying *completedness*, i.e., telicity, determining whether an event reaches a culmination or completion point at which a new state is introduced. Their approach describes each verb occurrence exclusively using features reflecting corpus-based statistics of the corresponding verb type. For each verb type, they collect the co-occurrence frequencies with 14 linguistic markers (e.g., present tense, perfect, combination with temporal adverbs) in an automatically parsed background corpus. They call these features *linguistic indicators* and train a variety of machine learning models based on 300 clauses, of which roughly 2/3 are *culminated*, i.e., telic. Their test set also contains about 300 clauses, corresponding to 204 distinct non-stative verbs. Their data sets are not available, but as this work is the most closely related to ours, we reimplement their approach and compare to it in Section 5.

Samardžić and Merlo (2016) create a model for real-world duration of events (as *short* or *long*) of English verbs as annotated in TimeBank (Pustejovsky et al., 2003). The model is informed by temporal boundedness information collected from parallel English-Serbian data. Their only features are how often the respective verb type was aligned to Serbian verbs carrying certain affixes that indicate perfectiveness or imperfectiveness. Their usage of "verb type" differs from ours as they do not lemmatize, i.e., they always predict that "falling" is a long event, while "fall" is short. Our approach shares the idea of projecting aspectual information from Slavic languages to English, but in contrast to classifying verb types, we classify whether an event type introduced by the verb constellation of a clause is telic or atelic, making use of a machine-readable dictionary for Czech instead of relying on affix information.

Loáiciga and Grisot (2016) create an automatic classifier for boundedness, defined as whether the endpoint of an event has occurred or not, and show that this is useful for picking the correct tense in French translations of the English Simple Past. Their classifier employs a similar but smaller feature set compared to ours. Other related work on predicting aspect include systems aiming at identifying lexical aspect (Siegel and McKeown, 2000; Friedrich and Palmer, 2014) or habituals (Mathew and Katz, 2009; Friedrich and Pinkal, 2015).

Cross-linguistic annotation projection approaches mostly make use of existing manually created annotations in the source language; similar to our approach, Diab and Resnik (2002) and Marasović et al. (2016) leverage properties of the source language to automatically induce annotations on the target side.

3 Data sets and annotation projection

We conduct our experiments based on two data sets: (a) English texts from MASC manually annotated for telicity, on which we train and test our computational models, and (b) a silver standard automatically extracted via annotation projection from the Czech-English part of the parallel corpus InterCorp, which we use as additional training data in order to improve our models.[2]

3.1 Gold standard: MASC (EN)

We create a new data set consisting of 10 English texts taken from MASC (Ide et al., 2010), annotated for telicity. Texts include two essays, a journal article, two blog texts, two history texts from travel guides, and three texts from the fic-

[1] In Czech, aspectual verb pairs may be related by affixes as in this example, but this is not always the case. They may even use different lexemes (Vintr, 2001).

[2] Annotations, guidelines and code available from https://github.com/annefried/telicity

	MASC (gold standard)	InterCorp (silver standard)	intersection
clauses (instances)	1863	457,000	-
% telic	82	55	-
% atelic	18	45	-
distinct verb types (lemmas)	567	2262	510
ambiguous verb types	70	1130	69

Table 1: Corpus statistics.

tion genre. Annotation was performed using the web-based SWAN system (Gühring et al., 2016). Annotators were given a short written manual with instructions. We model telicity for *dynamic* (*eventive*) verb occurrences because *stative* verbs (e.g., "like") do not have built-in endpoints by definition. Annotators choose one of the labels *telic* and *atelic* or they skip clauses that they consider to be stative. In a first round, each verb occurrence was labeled by three annotators (the second author of this paper plus two paid student assistants). They unanimously agreed on telicity labels for 1166 verb occurrences; these are directly used for the gold standard. Cases in which only two annotators agreed on a telicity label (the third annotator may have either disagreed or skipped the clause) are labeled by a fourth independent annotator (the first author), who did not have access to the labels of the first rounds. This second annotation round resulted in 697 further cases in which three annotators gave the same telicity label. Statistics for our final gold standard, which consists of all instances for which at least three out of the four annotators agreed, are shown in Table 1; "ambiguous" verb types are those for which the gold standard contains both telic and atelic instances. 510 of the 567 verb types also occur in the InterCorp silver standard, which provides training instances for 69 out of the 70 ambiguous verb types.

Finally, there are 446 cases for which no three annotators supplied the same label. Disagreement and skipping was mainly observed for verbs indicating attributions ("critics claim" or "the film uses"), which can be perceived either as statives or as instances of historic present. Other difficult cases include degree verbs ("increase"), aspectual verbs ("begin"), perception verbs ("hear"), iteratives ("flash") and the verb "do." For these cases, decisions how to treat them may have to be made depending on the concrete application; for now, they are excluded from our gold standard. Another source of error is that despite the training, annotators sometimes conflate their world knowledge (i.e., that some events necessarily come to an end eventually, such as the "swimming in the lake" in (2)) with the annotation task of determining telicity at a linguistic level.

3.2 Silver standard: InterCorp (EN-CZ)

We create a silver standard of approximately 457,000 labeled English verb occurrences (i.e., clauses) extracted from the InterCorp parallel corpus project (Čermák and Rosen, 2012). We leverage the sentence alignments, as well as part-of-speech and lemma information provided by InterCorp. We use the data from 151 sentence-aligned books (novels) of the Czech-English part of the corpus and further align the verbs of all 1:1-aligned sentence pairs to each other using the verbs' lemmas, achieving high precision by making sure that the translation of the verbs is licensed by the free online dictionary Glosbe.[3] We then look up the aspect of the Czech verb in Vallex 2.8.3 (Žabokrtský and Lopatková, 2007), a valency lexicon for Czech verbs, and project the label *telic* to English verb occurrences corresponding to a *perfective* Czech verb and the label *atelic* to instances translated using *imperfective* verbs.

Our annotation projection approach leverages the fact that most perfective Czech verbs will be translated into English using verb constellations that induce a telic event structure, as they describe one-time finished actions. Imperfective verbs, in contrast, are used for actions that are presented as unfinished, repeated or extending in time (Vintr, 2001). They are often, but not always, translated using atelic verb constellations. A notable exception is the English Progressive: "John was reading a book" signals an ongoing event in the past, which is telic at the word-sense level but would require translation using the imperfective Czech verb "četl." The initial corpus contained 4% sentences in the Progressive, out of which 89% were translated using imperfectives.[4] Due to the above

[3] https://glosbe.com

[4] For comparison, in the manually annotated validation

described mismatch, we remove all English Progressive sentences from our silver standard. Statistics for the final automatically created silver standard are shown in Table 1.

For validation, we sample 2402 instances from the above created silver standard and have our three annotators from the first annotation round mark them in the same way as the MASC data. Sampling picked one instance per verb type but was otherwise random. A majority agreement among the three annotators can be reached in 2126 cases (due to allowing skipping).[5] In this sample, 77.8% of the instances received the label telic from the human annotators, 61.5% received the label telic from the projection method. The accuracy of our projection method can be estimated as about 78%; F_1 for the telic class is 0.84, F_1 for atelic is 0.65. Errors made by the projection include for instance habituals, which use the imperfective in Czech but are not necessarily atelic at the event type level as in "John cycles to work every day."

4 Computational modeling

In this section, we describe the computational models for telicity classification, which we test on the MASC data and which we improve by adding the InterCorp silver standard data.

Features. We model each instance by means of a variety of syntactic-semantic features, using the toolkit provided by Friedrich et al. (2016).[6] Preprocessing is done using Stanford CoreNLP (Chen and Manning, 2014) based on dkpro (Eckart de Castilho and Gurevych, 2014). For the verb's lemma, the features include the WordNet (Fellbaum, 1998) sense and supersense and linguistic indicators (Siegel and McKeown, 2000) extracted from GigaWord (Graff et al., 2003). Using *only* the latter as features corresponds to the system by SMK2000 as described in Section 2. The feature set also describes the verb's subject and objects; among others their number, person, countability[7], their most frequent WordNet sense and the respective supersenses, and dependency relations between the argument and its governor(s). In addition, tense, voice and whether the clause is in the Perfect or Progressive aspect is reflected, as well as the presence of clausal (e.g., temporal) modifiers. For replicability we make the configuration files for the feature set available.

Classifier. We train L1-regularized multi-class logistic regression models using LIBLINEAR (Fan et al., 2008) with parameter settings ε=0.01 and bias=1. For each instance described by feature vector $\vec{x}$, the probability of each possible label y (here telic or atelic) is computed according to

$$P(y|\vec{x}) = \frac{1}{Z(\vec{x})} exp\left(\sum_{i=1}^{m} \lambda_i f_i(y, \vec{x})\right),$$

where f_i are the feature functions, λ_i are the weights learned for each feature function, and $Z(\vec{x})$ is a normalization constant (Klinger and Tomanek, 2007). The feature functions f_i indicate whether a particular feature is present, e.g., whether the tense of the verb is "past."

5 Experiments

Experimental settings. We evaluate our models via 10-fold cross validation (CV) on the MASC data set. We split the data into folds by documents in order to make sure that no training data from the same document is available for each instance in order to avoid an unfair bias. We report results in terms of accuracy, F_1 per class and macro-average F_1 (the harmonic mean of macro-average precision and recall). We test significance between differences in F_1 (for each class) using approximate randomization (Yeh, 2000; Padó, 2006) with $p < 0.1$ and significance between differences in accuracy using McNemar's test (McNemar, 1947) with $p < 0.01$. Table 2 shows our results: significantly different scores are marked with the same symbol where relevant (per column).

Results. A simple baseline of labeling each instance with the overall majority class (telic) has a very high accuracy, but the output of this baseline is uninformative and results in a low F_1. Rows titled "verb type" use the verb's lemma as their single feature and thus correspond to the informed baseline of using the training set majority class for each verb type. Rows labeled "+IC" indicate that the full set of instances with projected labels extracted from InterCorp has been added as additional training data in each fold; in rows titled "+ICs," the telic instances in InterCorp have been upsampled to match the 80:20 distribution in MASC. Our model using the full set of features significantly outperforms the verb type baseline

sample only 66% of Progressives received the label atelic.

[5] Of the 2402 cases, annotators completely agreed on 1577 cases (1114 telic, 203 atelic, 260 skipped). 85 cases were 2x atelic + 1x skipped, 219 cases were 2x telic + 1x skipped.

[6] https://github.com/annefried/sitent

[7] http://celex.mpi.nl

as well as SMK2000 (see † ‡ ∗). Using the additional training data from InterCorp results in a large improvement in the case of the difficult (because infrequent) atelic class (see ⋆), leading to the best overall results in terms of F_1. The best results regarding accuracy *and* F_1 are reached using the sampled version of the silver standard; the differences compared to the respective best scores in each column (in bold) are not significant.

Ablation experiments on the MASC data show that features describing the clause's main verb are most important: when ablating part-of-speech tag and tense and aspect (Progressive or Perfect), performance deteriorates by 1.8% in accuracy and 5% F_1, hinting at a correlation between telicity and choice of tense-aspect form. Whether this is due to an actual correlation of how telic and atelic verbs are used in context or merely due to annotation errors remains to be investigated in future work.

In sum, our experiments show that using annotations projected onto English text from parallel Czech text as cheap additional training data is a step forward to creating better models for the task of classifying telicity of verb occurrences.

6 Conclusion

Our model using a diverse set of features representing both verb-type relevant information and the context in which a verb occurs strongly outperformed previous work on predicting telicity (Siegel and McKeown, 2000). We have shown that silver standard data induced from parallel Czech-English data is useful for creating computational models for recognizing telicity in English. Our new manually annotated MASC data set is freely available; the projected annotations for InterCorp are published in a stand-off format due to license restrictions.

7 Future work

Aspectual distinctions made by one language rarely *completely* correspond to a linguistic phenomenon observed in another language. As we have discussed in Section 3.2, telicity in English and perfectiveness in Czech are closely related. As shown by our experiments, the projected labels cover useful information for the telicity classification task. One idea for future work is thus to leverage additional projected annotations from similar phenomena in additional languages, possibly improving overall performance by combining complementary information. Clustering more than two languages may also enable us to induce clusters corresponding to the different usages of imperfective verbs in Czech.

	Acc.	**F_1**	**F_1(telic)**	**F_1(atelic)**
maj. class	82.0	45.0	90.1	0.0
SMK2000	†83.0	63.9	†90.4	†26.8
SMK2000+IC	78.6	65.6	86.8	44.2
SMK2000+ICs	∗81.8	58.2	89.9	∗12.4
verb type	‡83.8	66.7	‡91.0	‡24.9
verb type+IC	82.4	73.5	89.0	57.1
verb type+ICs	85.1	72.2	91.2	∗51.9
our model	†‡**86.7**	74.5	†‡**92.2**	† ‡ ⋆53.7
our model+IC	82.3	**76.4**	88.6	**61.4**
our model+ICs	∗**86.2**	**76.2**	**91.6**	⋆∗**60.6**

Table 2: Results for telicity classification on MASC data (1863 instances), 10-fold CV.

The presence of endpoints has consequences for the temporal interpretation of a discourse (Smith, 1997; Smith and Erbaugh, 2005), as endpoints introduce new states and therefore signal an advancement of time. In English, boundedness, i.e., whether an endpoint of an event has actually occurred, is primarily signaled by the choice of tense and Progressive or Perfect aspect. In tense-less languages such as Mandarin Chinese, boundedness is a covert category and closely related to telicity. We plan to leverage similar ideas as presented in this paper to create temporal discourse parsing models for such languages.

When translating, telic and atelic constructions also require different lexical choices and appropriate selection of aspectual markers. Hence, telicity recognition is also relevant for machine translation research and could be a useful component in computer aided language learning systems, helping learners to select appropriate aspectual forms.

Acknowledgments

We thank Klára Jágrová, Irina Stenger and Andrea Fischer for their help with Czech-specific questions and with finding appropriate corpora, and Lucie Poláková for pointing us to Vallex. Melissa Peate Sørensen and Christine Bocionek helped with the annotation. Finally, thanks to the anonymous reviewers, Alexis Palmer, Manfred Pinkal, Andrea Horbach, Benjamin Roth, Katharina Kann and Heike Adel for their helpful comments. This research was supported in part by the MMCI Cluster of Excellence of the DFG.

References

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3):411–427.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 740–750.

Bernard Comrie. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2 of *Cambridge Textbooks in Linguistics*. Cambridge University Press.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 255–262.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland, pages 1–11.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachussetts.

Hana Filip. 1994. Aspect and the semantics of noun phrases. *Tense and aspect in discourse* 75:227.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, USA.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English Gigaword. *Linguistic Data Consortium, Philadelphia* .

Timo Gühring, Nicklas Linz, Rafael Theis, and Annemarie Friedrich. 2016. SWAN: an easy-to-use web-based annotation system. In *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*. Bochum, Germany.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden, pages 68–73.

Roman Klinger and Katrin Tomanek. 2007. Classical probabilistic models and conditional random fields. *TU Dortmund Algorithm Engineering Report* .

Sharid Loáiciga and Cristina Grisot. 2016. Predicting and using a pragmatic component of lexical aspect. *Linguistic Issues in Language Technology, Special issue on Modality in Natural Language Understanding* 13.

Ana Marasović, Mengfei Zhou, Alexis Palmer, and Anette Frank. 2016. Modal Sense Classification At Large: Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations. In *Linguistic Issues in Language Technology, Special issue on Modality in Natural Language Understanding*. CSLI Publications, Stanford, CA., volume 14 (2).

Thomas A. Mathew and E. Graham Katz. 2009. Supervised Categorization of Habitual and Episodic Sentences. In *Sixth Midwest Computational Linguistics Colloquium*. Bloomington, Indiana: Indiana University.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational linguistics* 14(2):15–28.

Sebastian Padó. 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. 2003. The TimeBank corpus. In *Corpus linguistics*. page 40.

Tanja Samardžić and Paola Merlo. 2016. Aspect-based learning of event duration using parallel corpora. In *Essays in Lexical Semantics and Computational Lexicography – In Honor of Adam Kilgarriff*, Springer Series Text, Speech, and Language Technology.

Eric V Siegel and Kathleen R McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics* 26(4):595–628.

Carlota S Smith. 1997. *The parameter of aspect*, volume 43. Springer Science & Business Media.

Carlota S Smith and Mary S Erbaugh. 2005. Temporal interpretation in Mandarin Chinese. *Linguistics* 43(4):713–756.

Josef Vintr. 2001. *Das Tschechische: Hauptzüge seiner Sprachstruktur in Gegenwart und Geschichte*. Sagner.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 947–953.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics* (87):41–60.