Automatic TIMEX2 Tagging of Korean News

SEOK BAE JANG, JENNIFER BALDWIN, AND INDERJEET MANI
Georgetown University

This article reports on a temporal tagger for Korean based on a Korean extension of the TIDES TIMEX2 guidelines. The extension, which primarily addresses the idiosyncrasies of Korean morphology, shows high inter-annotator reliability (0.893 F-measure for tag extent) when applied to a corpus of Korean newspaper articles. A machine-learning approach based on rote learning from a human-edited, automatically-derived dictionary of temporal expressions is compared with a second approach that adds manual patterns, and a third onethat tries to learn the patterns. Results for the first two are promising (0.87 F-measure for tag extent). Overall, the article shows that rote learning approaches can be very useful when language-specific features such as morphology are taken into account.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing - *Text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - *Linguistic processing*

General Terms: Languages

Additional Key Words and Phrases: Time, temporal information, temporal expressions, Korean

1. INTRODUCTION

Processing temporal information in natural language is valuable in question-answering (e.g., answering “when” questions by temporally anchoring events), information extraction (e.g., normalizing information for database entry), and summarization (temporally ordering information). This article reports on an approach to temporal tagging for Korean that uses a Korean extension to the TIDES TIMEX2 guidelines [Ferro et al. 2001] for annotating temporal information in natural language.

The US government’s DARPA TIDES program developed the TIMEX2 annotation scheme for marking the extent of English time expressions (with TIMEX2 tags) and normalizing their values. The scheme represents both absolute temporal expressions (e.g., “June 1, 2005”), as well as relative expressions that depend on context for their interpretation (e.g., “June 1,” “yesterday,” etc.). The TIMEX2 scheme has been adopted in part by the government in the Automatic Content Extraction (ACE) program's Relation Detection and Characterization (RDC) task, and in two Advanced Research and Development Activity (ARDA) TimeML summer workshops (timeml.org), where it has been extended in various ways. The ACE program led to a revised specification of TIMEX2 [Ferro et al. 2003] and the ARDA workshops to a revised specification in the form of TIMEX3 [Pustejovsky et al. 2004].

Author's address: Department of Linguistics, Georgetown University, 17th and O Streets, NW, Washington, DC 20057

There has been some initial work on extending TIMEX2 to other languages. A small parallel corpus of 95 Spanish-English dialogs has been annotated with TIMEX2 by a single bilingual annotator, based on tagging the English side and adapting it to the Spanish (available at timex2.mitre.org). There has also been some initial work on a set of Chinese guidelines [Gerber et al. 2004], as well as the development of a preliminary Hindi tagger for the TIDES Surprise Language experiment (see timex2.mitre.org). There have also been other automatic taggers reported for specific languages, e.g., Mani and Wilson [2000] and Setzer and Gaizauskas [2001] for English and Schilder and Habel [2001] for German. However, at this point there has been no systematic effort to evaluate how successful the annotation scheme's extension to another language is in terms of inter-annotator reliability or to develop and evaluate machine-learning approaches to temporal annotation that can be applied to multiple languages.

We begin by describing our Korean extensions to the guidelines in Section 2, based on manually annotating a corpus composed of 200 Korean newspaper articles excerpted from the corpus of the 21st Century Sejong Project.[1] In Section 3, we describe reliability across multiple human annotators. In Section 4, we present a machine-learning approach based on rote learning from a human-edited, automatically-derived dictionary of temporal expressions, followed, in Section 5, by its evaluation. Rote learning is the simplest of all learning methods, and it can also be the most efficient. It is attractive when a large sample of training data is available, but since annotated training data is hard to acquire, we report (in Section 6) on a method that augments the rote approach with manual patterns. In Section 7, we compare a third approach that tries to learn the patterns.

Although our work, which began in 2002, uses the 2001 TIMEX2 scheme, we discuss in Section 8 the implications based on the above-mentioned TIMEX2 and TIMEX3 revisions. In Section 8, we also discuss the implications for the 2004 Time Expression Recognition and Normalization (TERN) evaluation, sponsored by the US government.

2. APPLYING TIMEX2 TO KOREAN

2.1 Korean Extensions

In a temporal annotation scheme like TIMEX2, as Wilson et al. [2001] point out, the rules for extent are likely to be language-specific, whereas the rules for specifying normalized values are expected to be largely language-independent. The reason for the former is due to morphological and syntactic properties unique to each language; the reason for the latter is that the semantic representation for time being annotated is, by design, intended to be language-independent and used as an input to other computational inference procedures. Rather than developing a detailed guidelines document for Korean, as Gerber et al. [2004] do in their supplement for Chinese TIMEX2, our approach here is narrower in scope, and has resulted in a relatively short document specifying a set of extensions to the English guidelines. Here we summarize these extensions, followed in Section 2.2, by a corpus analysis.

[1] The 21st Century Sejong Project is a Korean national corpus project from 1998 sponsored by the Korean Minis try of Culture and Tourism. (www.sejong.or.kr). The sources of newspaper articles used in this article are the *C hosun Ilbo* and the *Hankook Ilbo*.

Korean is an agglutinative language; this has implications for the rules for extent. The English TIMEX2 guidelines state that temporal prepositions like “from” and “to” (as in “from 6 p.m. to 9 p.m.”) are not part of the extent. Since Korean instead uses postpositions that are bound morphemes (e.g., 부터 “from” and 까지 “to”), one faces a choice of either including postpositions in the extent, or allowing subword TIMEX2 tags that exclude the postpositions. We made the latter choice, assuming it is more conservative. Thus, we have:

전시회가 9 월 7 일부터 20 일까지 열린다.
Exhibition-NOM September 7th-FROM 20th-TO be_held-FUTURE.
The exhibition will be held from September 7th to 20th.
전시회가 <TIMEX2 VAL=“2000-09-07”>9 월 7 일</TIMEX2>부터 <TIMEX2 VAL=“2000-09-20”>20 일</TIMEX2>까지 열린다.

Another aspect of the agglutinative morphology concerns suffixes that have temporal meanings, like “간(during)” or “쯤 (about)”:

회의가 3 일간만 열린다.
Meeting-NOM 3 day-Suffix(Duration)-ONLY be_held-FUTURE
The meeting will be held for 3 days.
회의가 <TIMEX2 VAL=“P3D”>3 일간</TIMEX2>만 열린다.

회의가 9 월쯤에 열린다.
Meeting-NOM September-Suffix(About)-IN be_held-FUTURE
The meeting will be held in about September.
회의가 <TIMEX2 VAL=“2000-09” MOD=“APPROX”>9 월쯤</TIMEX2>에 열린다.

In addition, the English guidelines require vague conjoined expressions like “two or three hours” to be annotated with two tags, whereas “two or three” is a single word 두세 in Korean. This phenomenon can occur systematically in Korean temporal expressions that have number morphemes:

두세 시간 / 열두세 시간 / 백열두세 시간 …
two-three hour / ten-two-three hour / hundred-ten-two-three hour
2 or 3 hours / 12 or 13 hours / 112 or 113 hours

Due to the non-fusional characteristic of Korean morphology, the separation of words into their component morphemes is straightforward; hence we annotate each morpheme separately:

옷을 비눗물에 담그고 두세 시간을 둔다.
Clothes-ACC soapy_water-IN soak-CONJ 2 or 3 hour-ACC keep-PRESENT.
Soak clothes in soapy water for 2 or 3 hours.
옷을 비눗물에 담그고 <TIMEX2 VAL=“PT2H”>두</TIMEX2><TIMEX2 VAL=“PT3H”>세 시간</TIMEX2>을 둔다.

Table I. Distribution of Different Varieties of Temporal Expressions

	English (105,911 words)			Korean (30,887 words)		
	Occurrence	Per Word	%	Occurrence	Per Word	%
Absolute	422	0.0040	25.8%	160	0.0052	18.1%
Relative	699	0.0066	42.8%	489	0.0158	55.3%
Duration	274	0.0026	16.8%	173	0.0056	19.5%
REF	239	0.0023	14.6%	63	0.0020	7.1%
Total	1,634	0.0154	100.0%	885	0.0287	100.0%

Apart from this, however, the annotation scheme carried over very well. Note that the TIMEX2 guidelines represent time values only in the Gregorian calendar. Although Korean does make use of alternative calendars (namely, lunar calendar and the Tangun era calendar), no references to them were observed in the corpus.

2.2 Corpus Analysis

Table I shows a quantitative comparison of temporal expressions in English and Korean news articles. The English articles are from *The New York Times*, and the Korean articles are from the corpus of the 21st century Sejong Project. The English corpus is composed of 100 articles. We show the counts for this corpus, and for purposes of comparison, we also show the corresponding counts for a 100-article randomly chosen subcorpus of the Korean 200-article corpus. The discrepancy between the number of words in the two languages is because English articles are much longer on the average than Korean ones; in addition, postpositions in Korean, which are bound morphemes, are not counted as separate words.

It can be seen from Table I that the various types of time expressions are ordered in a similar way in Korean and English, with relative expressions dominating. Also, the percentage of durational expression is quite similar. This means that automatic taggers for Korean will encounter these two categories of temporal expressions roughly as often as in English. However, the percentages for the other categories[2] are different. Further research is required to test if the latter differences reflect particular styles of news writing.

3. INTER-ANNOTATOR AGREEMENT

To establish how consistently humans can perform TIMEX2 annotation in Korean, we evaluated inter-annotator reliability involving two human annotators. The annotators are (A) the first author of this article, who is experienced in TIMEX2 annotation, and a second annotator (B), based in Korea, who works in corpus linguistics but had no prior experience with temporal tagging. The test data set used in building inter-annotator scores consists of 30 documents extracted from the training data. Annotator B was given the guidelines for English, a description of the extensions for Korean, and a separate

[2] REF stands for TIMEX2 "referential" expressions whose values are expressed by the primitive tokens PAST, PRESENT or FUTURE, e.g., <TIMEX2 VAL="PAST_REF">lately</TIMEX2>.

Table II. Inter-Annotator Agreement on 30 Documents

	POS	ACT	CORR	Precision	Recall	F-measure
Extent	227	223	201	.910	.885	.893
Value	200	195	181	.928	.905	.916
Granularity	5	0	0	0	0	0
Mod	6	5	4	.800	.667	.727
Non-Specific	0	1	0	0	0	0
Periodicity	5	10	5	.500	1.000	.667
Set	5	10	5	.500	1.000	.667

sample of tagged TIMEX2 Korean documents. B then annotated the 30 test documents without further communication with A.

Table II shows the F-measure inter-annotator results returned by ScoreV3, a TIMEX2 scorer (timex2.mitre.org). POS (for Possible tags) refers to a reference annotation while ACT (Actual tags) refers to a candidate annotation. CORR (Correct) indicates the number of instances judged correct by the scorer. Here we arbitrarily choose one of the annotators (A) as the reference and the other (B) as the candidate; swapping them merely swaps Precision and Recall, without affecting the F-measure.

The disagreements mainly come from annotator B skipping Korean durational suffixes and spuriously annotating non-trigger words (i.e., non-trigger adverbs). There are also some mismatches in the interpretation of temporal meanings. For example "하루(one day)" is mistakenly annotated by B as a set-denoting expression instead of a durational expression. The scores on dimensions, other than Extent and Value, are on a very small number of examples, and as such should be discounted.

In other work, inter-annotator reliability on English has been reported at 0.79 F-measure for extent and 0.86 F-measure for values [Mani 2004]. However, the latter result involves 5 annotators on 193 documents; thus the different scales of the evaluations make them difficult to compare.

4. A BASIC ROTE LEARNING METHOD

We developed an automatic tagger called KTX (Korean Temporal eXpression tagger). The tagger uses a rote learning method based on inducing a dictionary from the training data. Figure 1 illustrates the control architecture of KTX.

KTX first semi-automatically builds a dictionary of temporal expressions from a training corpus. While the absolute temporal expressions are easy to induce automatically, the intervention of a human in building a dictionary is required for some relative expressions and durational expressions that are accompanied by temporal pre-modifiers, post-modifiers and suffixes. Given our earlier annotation decision to use sub-word tags, the precise temporal expressions have to be extracted from the words. If we extend KTX to use part-of-speech information and information about temporal modifiers (which would in turn require a full morphological analyzer and additional modifier dictionaries), this step may be fully automated.

KTX then flags temporal expressions in the input document using the dictionary. The base date is taken from the publication date and time stated in the input document (found in an XML header in each document). KTX then carries out simple morphological analysis to find temporal expressions based on a stop word list that includes simple and complex postpositions and punctuation, etc. This morphological analysis is very simple, but fortunately there are few problems of morphological disambiguation in temporal expressions because KTX already has the temporal word dictionary. This step enables us to get relatively fast and automatic temporal annotation without many morphological ambiguities.

Most absolute temporal expressions in Korean can be used as durations when accompanied by durational modifiers or suffixes:

회의가 3 일에 열린다.

Meeting-NOM 3rd-ON be_held-FUTURE
The meeting will be held on the 3rd .

회의가 <TIMEX2 VAL="2000-06-03">3 일</TIMEX2>에 열린다.

회의가 3 일간 열린다.

Meeting-NOM 3days- Suffix(Duration) be_held-FUTURE
The meeting will be held for 3 days.

회의가 <TIMEX2 VAL="P3D">3 일간</TIMEX2> 열린다.

KTX therefore uses a module that disambiguates absolute and relative temporal expressions from durations. This module is based on several heuristic rules obtained by inspection of the training corpus. These rules are arranged so that the next rule is tried if the previous one fails, with a rule's success leading to the expression being classified as a duration: temporal expression followed by a word (usually expressed as a noun) having a durational meaning; temporal expression preceded by a word that has a durational meaning; and temporal expression with a suffix having a durational meaning.

The three rules above are used with 8 nouns as pre-modifiers, 17 nouns as post-modifiers, and 7 suffixes. In addition to base forms, some fixed patterns (e.g., noun + postposition, noun + suffix) of those words are also used, since it is difficult to recognize exact modifiers from modified temporal strings without part-of-speech information.

Table III shows a sample of the temporal expression dictionary generated semi-automatically from training data. In the Item column, the first sub-column contains the actual temporal expressions and the second sub-column contains the (semi-colon delimited) temporal attributes for the expression. The first field in the temporal attributes indicates the type of temporal expression (Y, M, D, W, H, MN, S, R). The second field shows whether the expression is an absolute date, a relative date, a duration, or a

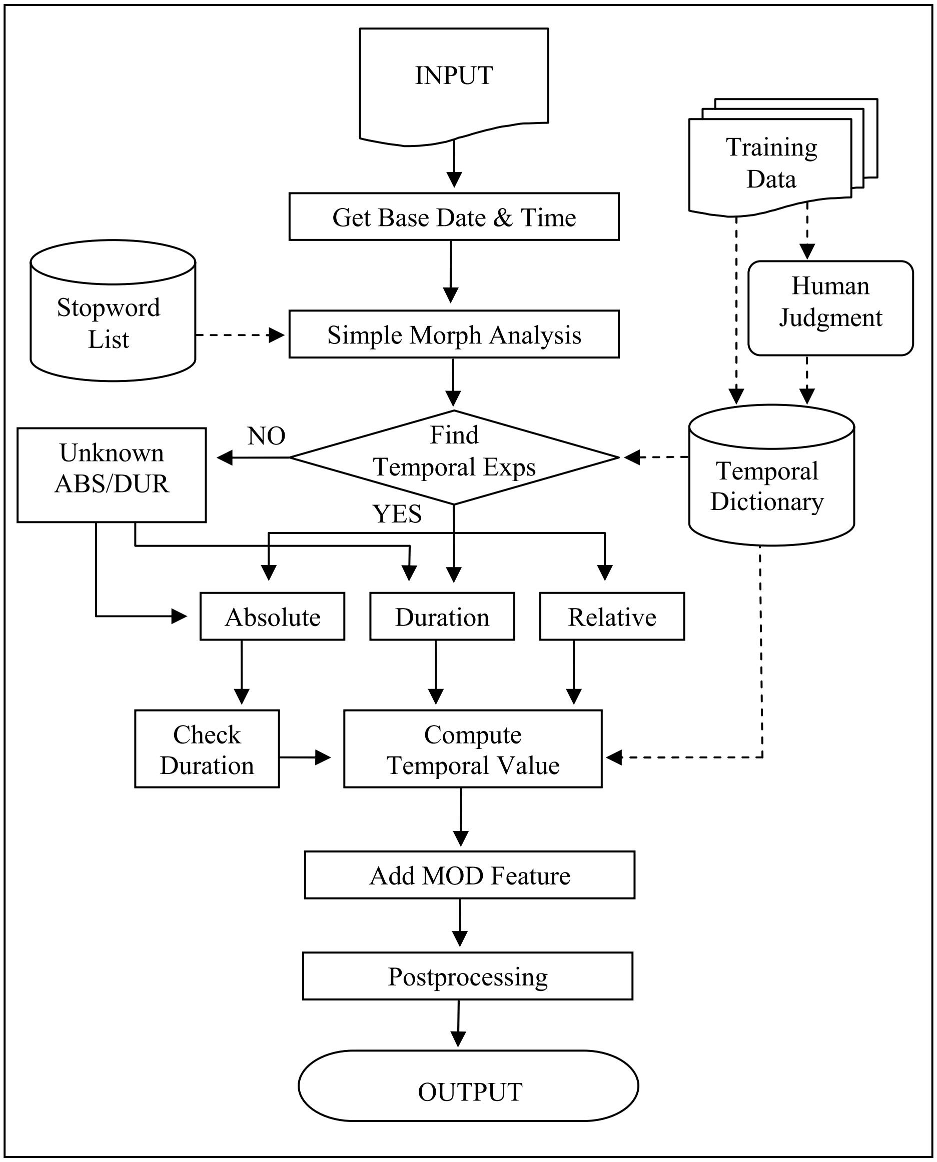


Fig. 1. KTX system architecture.

referential (REF) expression. The last field is a temporal offset where the expression is relative and an actual temporal value when the expression is an absolute, durational, or referential expression. When the expression is a relative, the offset indicates the

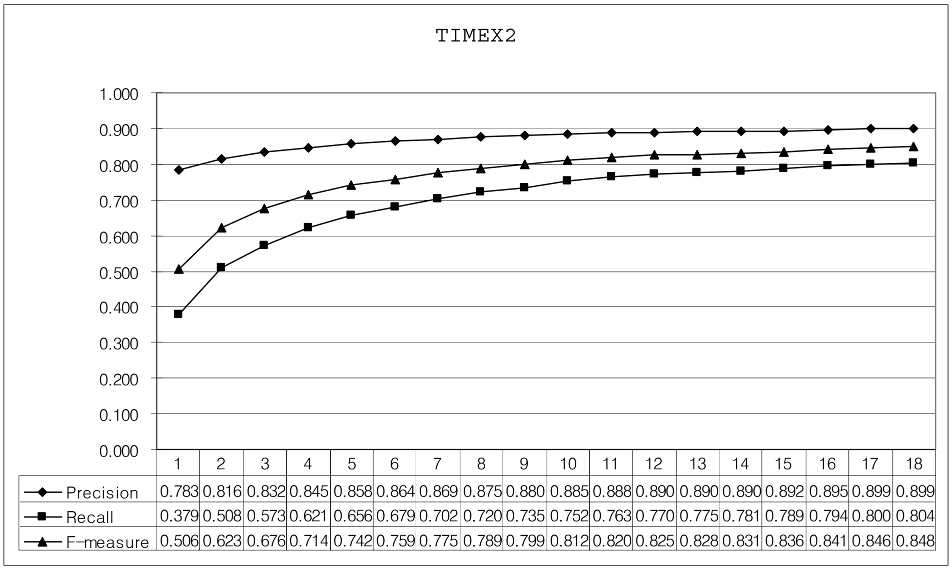
Table III. Sample of Temporal Expression Dictionary

Type	Item		Type	Item	
Absolute	1997 년 (year 1997)	Y;ABS;1997	Relative	지난달 (last month)	M;REL;-1
Absolute	21 세기 (21st century)	Y;ABS;20	Relative	오늘 (today)	D;REL;0
Absolute	5 월 (May)	M;ABS;05	Relative	다음주 (next week)	W;REL;-1
Absolute	20 일 (20 days)	D;ABS;20	Relative	모레 (the day after tomorrow)	D;REL;2
Absolute	겨울 (Winter)	M;ABS;WI	Duration	한달 (one month)	M;DUR;1
Absolute	하반기 (2nd half)	M;ABS;H2	Duration	10 개월 (for 10 months)	M;DUR;10
Absolute	3 분기 (3rd quarter)	M;ABS;Q3	Duration	사흘 (for 3 days)	D;DUR;3
Absolute	10 시 (10 hours)	H;ABS;10	Duration	한해 (for 1 year)	Y;DUR;1
Absolute	30 분 (30 minutes)	MN;ABS;30	REF	현재 (currently)	R;REF;PRESENT_REF
Absolute	29 초 (29 seconds)	S;ABS;29	REF	과거 (past)	R;REF;PAST_REF
Relative	내년 (next year)	Y;REL;1	REF	향후 (future)	R;REF;FUTURE_REF

difference between the temporal value of current expression and the base temporal value. For example, if the type of relative expression is D and the offset is -2, it means that the expression indicates two days before the base date.

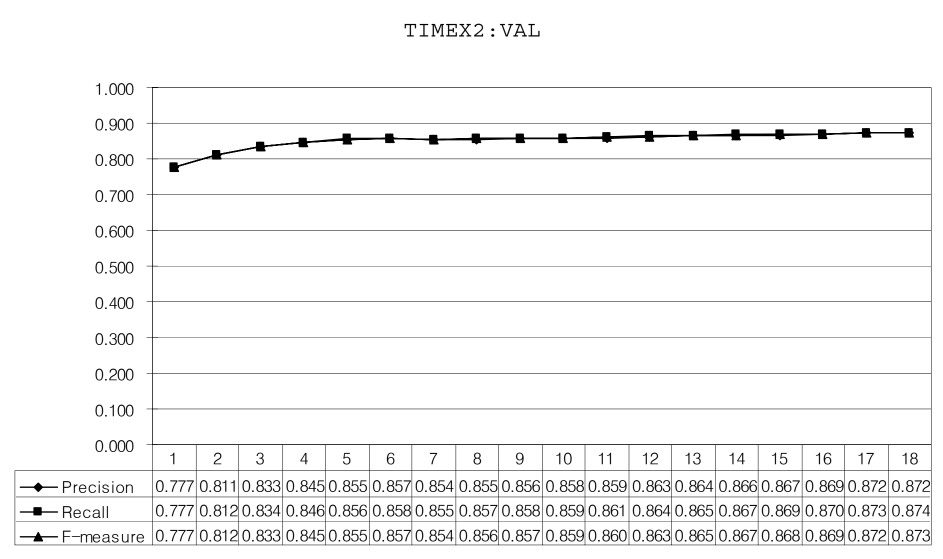
5. EFFECT OF TRAINING CORPUS SIZE

We now report on the influence of training set size on learner performance. Using a set of 200 documents, we set the test data size to 20 and varied the training data size from 10 to 180 in increments of 10. Each test set and training set was gathered in 20 folds in order to



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
◆ Precision	0.783	0.816	0.832	0.845	0.858	0.864	0.869	0.875	0.880	0.885	0.888	0.890	0.890	0.890	0.892	0.895	0.899	0.899
■ Recall	0.379	0.508	0.573	0.621	0.656	0.679	0.702	0.720	0.735	0.752	0.763	0.770	0.775	0.781	0.789	0.794	0.800	0.804
▲ F-measure	0.506	0.623	0.676	0.714	0.742	0.759	0.775	0.789	0.799	0.812	0.820	0.825	0.828	0.831	0.836	0.841	0.846	0.848

Fig. 2. KTX learning curve for TIMEX2 tags.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
◆ Precision	0.777	0.811	0.833	0.845	0.855	0.857	0.854	0.855	0.856	0.858	0.859	0.863	0.864	0.866	0.867	0.869	0.872	0.872
■ Recall	0.777	0.812	0.834	0.846	0.856	0.858	0.855	0.857	0.858	0.859	0.861	0.864	0.865	0.867	0.869	0.870	0.873	0.874
▲ F-measure	0.777	0.812	0.833	0.845	0.855	0.857	0.854	0.856	0.857	0.859	0.860	0.863	0.865	0.867	0.868	0.869	0.872	0.873

Fig. 3. KTX learning curve for TIMEX2 values.

minimize the effect of sampling of particular training data sets. The final results were acquired from the average scores of 20 folds. Each training set results in an induced dictionary, which is used against the test set.

The results are shown in Figures 2 and 3. From Figure 2, the F-measures of TIMEX2 tags are seen to increase continuously with the increase in training data size. The precision, which starts off high, is slightly improved, while recall goes up sharply, as might be expected. Figure 3 shows that the F-measures of time values start off high andincrease only slightly with additional training data, unlike the F-measures for extent. To increase to 0.90 F-measure and beyond, hand-created patterns may have to be added to handle contextual information for certain types of relative times.

6. AUGMENTING ROTE LEARNING WITH PATTERNS

The simple approach of only learning from the dictionary induced from the training data (manually post-edited as described in Section 4) is problematic when the training data set is small. We therefore compare the basic rote approach with one where patterns are added by a human.

The dictionary (“DICT-M” in the following discussion) is composed of the 170 temporal expressions that were semi-automatically derived from the training data of 100 documents. There are 121 absolute expressions: 50 for year, 16 for month, 26 for day, and 29 for time of day. There are also 16 relative expressions (year, month, day, and time of day), 28 expressions of duration (year, month, day, and time of day), and 5 REF expressions (past, present, and future).

DICT-M+H is DICT-M augmented with additional expressions, such as years to cover the 20th to 21st centuries, as well as everyday temporal expressions that were absent in the training data. In all, DICT-M+H comprises 460 temporal expressions that occurred in the training data of 100 documents as well as some commonly used expressions that were not present in the training corpus. DICT-M+H has 181 absolute expressions, 29 relative expressions, 188 expressions of time of day, 55 expressions for duration, and 7 REF expressions.

A large number of absolute temporal expressions in the test data are simply not present in the training data used to derive the dictionary semi-automatically. Fortunately, the morphological structure of absolute temporal expressions in Korean consists of a number and a word denoting a temporal counting unit. We developed a pattern-matching module to identify absolute temporal expressions based on this feature so as to identify absolute temporal expressions from input strings and to tag them without a temporal expression dictionary. An additional extension to KTX is an unknown absolute temporal word module (UNK-ABS) and an unknown durational word module (UNK-DUR) that identify absolute expressions and durations from unknown strings.

We now compare the performance of these extensions on a held-out 100-document test set distinct from the 100-document training set. The results are shown in Table IV. Here POS is the reference human annotation (from annotator A), whereas ACT is KTX (the candidate system) annotation. Note that KTX doesn’t identify tag attributes other than temporal extent or value. The reason the value F-measures can be higher than the extent F-measures is as follows. The scorer (version ScoreV3) flags occurrences of tags in a candidate annotation which do not occur in exactly the same position in the reference annotation as errors of extent. When the positions of the candidate and reference tags

Table IV. KTX Accuracy with Various Dictionaries

		POS	ACT	CORR	Precision	Recall	F-measure
DICT-M	Extent	877	654	566	.865	.645	.739
	Value	640	642	530	.826	.828	.827
DICT-M+H	Extent	877	801	716	.894	.816	.853
	Value	775	777	677	.871	.874	.872
DICT-M+ UNK-ABS/DUR	Extent	877	842	715	.849	.815	.832
	Value	798	800	667	.834	.836	.835
DICT-M+H+ UNK-ABS/DUR	Extent	877	864	757	.876	.863	.870
	Value	824	826	717	.868	.870	.869

coincide or overlap within a particular window size, the scorer compares the values of such expressions, scoring the values correct if the candidate and reference values are equivalent.

From Table IV, it can be seen that the recognition of temporal expressions is better when using a dictionary (DICT-M+H) extended beyond the training corpus and with modules (UNK-ABS/DUR) that identify absolute expressions and durations. The F-measure of extent is the highest (0.870) when we use DICT-M+H+UNK-ABS/DUR, but the F-measure of value is the highest (0.872) when we use DICT-M+H, because UNK modules over-generate temporal tags. The module UNK-ABS/DUR increases the recall of extent effectively, but the accuracy of extent drops slightly when guessing unknown temporal words.

7. INDUCTIVE LEARNING OF PATTERNS

While rote learning is simple and efficient, it is desirable for the learner to have some degree of generalization. UNK-ABS provides a degree of generalization, but this is based on hand-created patterns. We therefore wanted to learn patterns automatically. Previous machine-learning approaches such as that of Riloff [1996] addressed learning patterns from annotated data, but such approaches presuppose syntactic analysis of expressions in the target language. To develop an approach that requires minimal porting across languages, we developed another learner, TDL (Task-Driven Learner) [Baldwin 2002], based on the idea of a learner that is specific to tagging time expressions, but that isn't dependent on language-specific information. TDL uses an inductive learning approach to first build a lexicon of time expressions entirely automatically from the training data and then to analyze the mapping between strings and temporal tag attribute values. This automatic analysis is used to learn rules for tagging days of the week, months of the year, durations, etc. The overall architecture is shown in Figure 4 (for more details, see Baldwin [2002]).

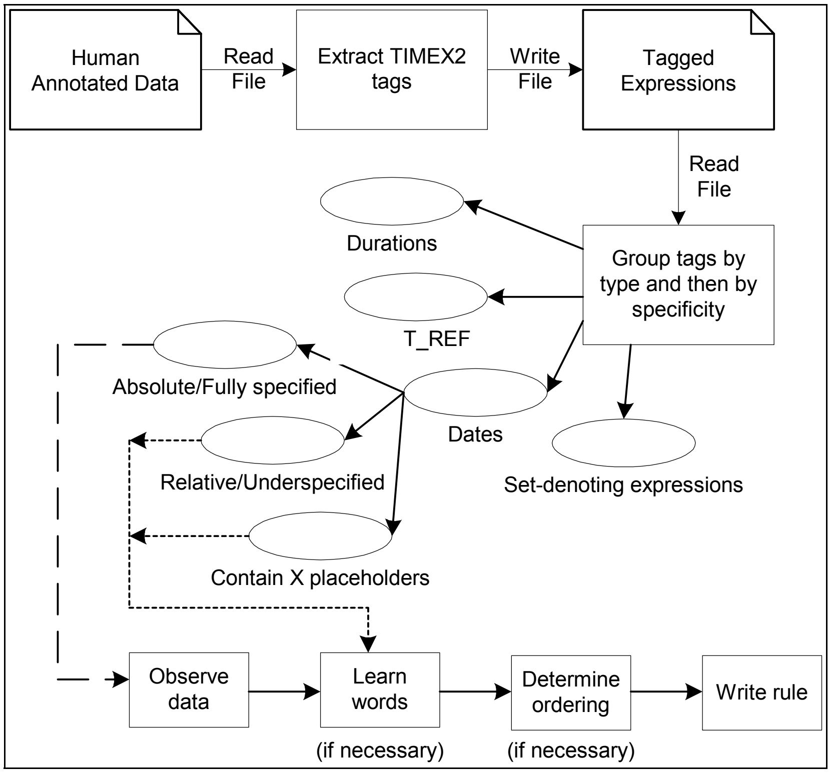


Fig. 4. TDL architecture.

In TDL, a time expression and its TIMEX2 tag information form a training example for learning the mappings between strings and the values of temporal attributes. For example, a collection of similar date examples like <TIMEX2 VAL="2001-02-17">February 17, 2001</TIMEX2> will generate a rule of the form:

Pattern([?M,?D,comma,?Y]) ⇒ Value(Year(?Y),Month(?M),DayofMonth(?D))

The rule is assigned a confidence based on the frequency of the pattern. Thus, given multiple Korean examples involving month expressions such as <TIMEX2 VAL="XXXX-09">9 월</TIMEX2>, TDL should learn that 9 월 is the expression for month 9, i.e., September. This leads to the induced rule:

Pattern([?N,월]) ⇒ Value(Month(?N))

For periods, given multiple examples such as <TIMEX2 VAL="P3M">3 개월 </TIMEX2>, the system should learn that 개월 can map to a month period, leading to the induced rule:

Pattern([?N,개,월]) ⇒ Value(MonthPeriod(?N))

Similar inductive learning occurs with other types of time expressions and for all other TIMEX2 attributes except MOD (the latter is left for future work).

Once rules are learned, they are all applied in the order they were learned. In order to resolve a test example that is a relative time, TDL uses information from the run-time context. Assume that TDL has learned that 토요일 maps to WXX-06, i.e., Saturday. Given the bare relative date 토요일, TDL finds dates for the 7-day window around the publication date using the calendar, finding the first Saturday within that window. Such heuristics would be improved if TDL could learn about dependencies on neighboring words corresponding to "next" or "last."

TDL knows about time and the structure of ISO temporal domains, but it does not rely on knowledge about Korean. It does make language-dependent assumptions about the extents of time expressions, such as assuming that month and day-of-week names are one "word" long. To date, we have used TDL as-is to serve as a baseline, to see how far we can get in a time-pattern learner that lacks language-specific information.

The performance of TDL on Korean is shown in Table V, along with data on English and French for comparison. The size of the training set is 122 documents for English, 30 documents for French, and 100 documents for Korean. We can see that the Korean performance is very low, due to the low recognition rate of Korean temporal expressions. TDL learned only certain fixed patterns, such as temporal expressions that consist of a number and a word denoting a temporal counter and some bare temporal words without modifiers, postpositions, and suffixes. Adding a Korean morphological analyzer would improve the performance considerably.

Table V. TDL Accuracy for English, French, and Korean

Language		Docs	POS	ACT	CORR	Precision	Recall	F-measure
English	Extent	71	1,204	933	602	.645	.500	.563
	Val	71	544	546	457	.837	.840	.839
French	Extent	15	204	160	110	.688	.539	.604
	Val	15	109	110	92	.836	.844	.840
Korean	Extent	100	577	488	117	.240	.203	.220
	Val	100	169	160	15	.094	.089	.091

8. RELATED ANNOTATION EFFORTS

The revised TIMEX2 guidelines [Ferro et al. 2003] include new primitive tokens for time values, standards for encoding time zones, representation of values for event-anchored time expressions under certain situations, and the inclusion of anchor-points in time and anchor directions in time values (e.g., for durations that are anchored at a start or end points). All these changes would require augmenting our Korean corpora accordingly; but such extensions fall in naturally with our work on human and machine annotation. The revisions also include clarifications of rules for what gets marked, which could again be propagated to our annotations.

The 2004 Time Expression Recognition and Normalization (TERN) evaluation (timex2.mitre.org) is aimed at TIMEX2 tagging based on the latest TIMEX2 guidelines for English and Chinese text. Given our comments about the updated English guidelines, our approach to Korean is clearly compatible. However, the Chinese guidelines [Gerber et al. 2004] are much broader in scope than our work on Korean guidelines in providing a detailed specification of guidelines for a language. Since Korean uses a writing system with word spacing (though the spacing rules are used inconsistently), whereas Chinese does not make use of word spacing, there are relatively fewer problems of extent marking in Korean compared to Chinese.

TimeML [Pustejovsky et al. 2004] is broader in scope than TIMEX2. It is concerned with annotating events, time expressions, and the anchoring of the former to the latter. TimeML includes revisions to TIMEX2, in the form of TIMEX3 tags. TIMEX3 uses a functional style of annotating time values, so that “last week” could be represented not only by the time value but also by an expression that could be evaluated to compute the value, namely, that it is the predecessor week of the week preceding the reference date. This sort of extension, being semantic in nature, extends naturally to Korean, as does the inclusion of event-dependent time expressions. Following Setzer and Gaizauskas [2001], TIMEX3 also annotates temporal prepositions with a SIGNAL tag, which fits naturally with Korean postpositions.

9. CONCLUSION

In this article we described a temporal annotation approach for Korean. We first extended temporal annotation guidelines developed for English to handle Korean news articles. The annotation scheme extended well, except for some changes made to accommodate the idiosyncrasies of Korean morphology. Inter-annotator agreement on extent and value was high (0.893 and 0.916 F-measures, respectively). We also implemented a Korean temporal tagger, called KTX, based on rote learning and then investigated the impact of extending it with manual augmentations to the dictionary and hand-created patterns, with a net result of F-measures of 0.87 on extent and 0.869 on value. In comparison, the English TIMEX2 tagger TempEx [Mani and Wilson 2000] based on hand-created patterns scores 0.76 F-measure for extent and 0.82 F-measure for time values [Mani 2004]. Another approach we investigated dispenses with human-created patterns or augmentations, using inductive learning to acquire patterns automatically, based on knowledge about temporal domains. TDL obtains an F-measure of 0.220 for extent and 0.091 for values; but the latter result was obtained without the benefit of Korean morphology.

Our experience with rote learning is quite encouraging. Using an automatically induced dictionary with minimum human editing gives us an 0.739 F-measure on extent and 0.827 F-measure on value, which is competitive with elaborate pattern-matching systems based on hand-created patterns like TempEx. Additional manual augmentation of the dictionary with hand-created patterns helps boost performance further. We conclude from this that rote methods are quite promising compared to using hand-created patterns, provided a certain degree of language-specific knowledge (in this case, knowledge of time-related Korean morphology) is given, in our case by humans. In contrast, a machine-learning approach like TDL that tries to learn patterns requires more language-specific knowledge. Here, our work provides a baseline value for such approaches.

The annotated corpora, as well as the KTX tagger, are available for download at timex2.mitre.org. In the future, we plan to extend TDL to use knowledge of Korean morphology and to use the dictionary format that KTX uses. We are also investigating other learning approaches that could combine co-training with KTX and other learners.

REFERENCES

BALDWIN, J. 2002. Instance-based machine learning approach for multilingual time expression annotation. MS thesis, Dept. of Linguistics, Georgetown Univ.

FERRO, L., MANI, I.., SUNDHEIM, B., AND WILSON, G. 2001. TIDES Temporal Annotation Guidelines Draft - Version 1.02. MITRE Tech. Rep. MTR 01W000004. The MITRE Corp., McLean, VA.

FERRO, L., GERBER, L., MANI, I., SUNDHEIM, B., AND WILSON, G. 2003. TIDES 2003 standard for the annotation of temporal expressions. Sept. 2003. timex2.mitre.org.

GERBER, L., HUANG, S., AND WANG, X. 2003. Standard for the annotation of temporal expressions, Chinese supplement draft, April 2004. timex2.mitre.org.

MANI, I. AND WILSON, G. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (New Brunswick, NJ).

MANI, I. 2004. Recent developments in temporal information extraction. In *Proceedings of the Conference on Recent Advances In Natural Language Processing*. N. Nicolov and K. Mitkov, eds. John Benjamins, to appear.

PUSTEJOVSKY, J., INGRIA, B., SAURI, R., CASTANO, J., LITTMAN, J., GAIZAUSKAS, R., SETZER, A., KATZ, G., AND MANI, I. 2004. The specification language TimeML. In *The Language of Time: A Reader*. I. Mani et al, eds., Oxford University Press, to appear.

RILOFF, E. 1996. An empirical study of automated dictionary construction for information extraction in three domains. *AI Journal*. *85* (Aug. 1996).

SCHILDER, F. AND HABEL, C. 2001. From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing* (EACL-ACL 2001, Toulouse, France, July 2001).

SETZER, A. AND GAIZAUSKAS, R. 2001. A pilot study on annotating temporal relations in text. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing* (EACL-ACL 2001, Toulouse, France, July 2001).

Wilson, G., Mani, I., Sundheim, B., and Ferro, L. 2001. A multilingual approach to annotating and extracting temporal information. In *Proceedings of the Workshop for Temporal and Spatial Information Processing* (EACL-ACL 2001,Toulouse, France, July 2001).

Received January 2004; accepted June 2004