УДК 81’32

РАЗРАБОТКА СИНТАКСИЧЕСКОГО, ЛЕКСИЧЕСКОГО И МОРФОЛОГИЧЕСКОГО НАБОРОВ МЕТОК ДЛЯ ГРАММАТИЧЕСКОЙ РАЗМЕТКИ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

© О.А.Макажанов, О.Е.Махамбетов, И.М.Сабыргалиев, Ж.А.Есенбаев

В статье описывается процесс разработки синтаксического, лексического и морфологического наборов меток для Казахского языка. Морфологическая разметка включает в себя набор меток для обозначения основных грамматических свойств имен и глаголов. Словообразующие суффиксы помечаются маркерами перехода между соответствующими частями речи. Для лексической разметки мы разработали гибкий набор меток, который, в зависимости от приложения, может содержать определенное их количество. Синтаксическая разметка представляет собой компактный набор меток, обозначающий стандартные синтаксические категории. Данная статья является расширенным вариантом статьи [1], опубликованной в сборнике трудов конференции «Theory, Engineering, Language» (TEL 2014). Помимо оригинального материала, статья включает в себя дополнительный раздел по методике работы с аннотаторами при создании разметок.

Ключевые слова: вычислительная лингвистика, разработка наборов грамматических меток, синтаксис, части речи, морфология.

1. Введение

Задачи автоматизации морфологического, лексического и синтаксического разборов являются базовыми в области обработки естественного языка. Их успешное решение позволяет решить либо улучшить качество решения целого ряда прикладных задач, таких как поиск по леммам, машинный перевод, извлечение из текста семантических связей и именных сущностей, и многих других. В свою очередь, для автоматизации разборов необходимо иметь соответствующие наборы меток, максимально отвечающие особенностям языка.

Мы столкнулись с проблемой отсутствия наборов грамматических разметок для казахского языка во время разметки данных для корпуса казахского языка [2]. Стоит отметить, что в работе А.Шарипбаева [3] формализованы процессы сегментации и генерации словоформ, однако отсутствует формальное описание морфологического и лексического наборов меток. В настоящей статье мы описываем методику создания подобных разметок для казахского языка.

При создании наборов разметок мы руководствовались мировым опытом [4-7], а также старались в максимальной степени учесть агглютинативную природу языка и его сложную морфологию. В основу созданного лексического набора меток легла позиционная система [5-6], в которой, кроме обозначения части речи, в метках закодированы различные грамматические категории. Разработанная позиционная разметка состоит из 36-и базовых меток; в зависимости от требуемой полноты разбора, позволяет учитывать до девяти грамматических признаков и может быть отображена в универсальную лексическую разметку [8]. Предварительные эксперименты со сравнительно небольшим набором тренировочных данных показали, что автоматические разметчики, основанные на методе последовательной разметки, работают точнее с разметками с меньшим количеством учитываемых грамматических признаков [1: 129].

2. Синтаксическая разметка

Таблица 1.

Синтаксическая разметка

№	Метка	Описание	Эквиваленты Penn Treebank
1	S	Простое предложение	S
2	BSS	Главное предложение	S
3	BGS	Зависимое предложение	SBAR, SBARQ
4	BAS	Подлежащее	NP
5	BND	Сказуемое	VP
6	TOL	Дополнение	NP, WHNP
7	ANT	Определение	ADJP
8	PYS	Обстоятельство	PP, WHP, ADVP, WHADVP
9	X	Пустой / неоднозначный член	X

Синтаксическая разметка описана в Таблице 1, содержащей наименование и описание меток, а также эквиваленты из широко употребляемого набора меток, Penn tagset [6]. Так как фразеоло-

гизмы часто составляют одну синтаксическую единицу, они также помечаются данной разметкой путем присвоения метке соответствующего бинарного атрибута. Приведенная разметка не приспособлена к построению многоуровневых деревьев разбора [6]. В настоящее время ведется работа по ее модификации.

3. Лексическая и морфологическая разметки

Разработанный набор морфологических обозначений учитывает основные грамматические свойства имен (число, принадлежность, падеж, лицо) и глаголов (залог, отрицание, наклонение, время, лицо) [1: 126]. Деривативные суффиксы обозначаются соответствующим переходом между частями речи.

Для лексической разметки используется расширенный вариант набора частей речи, в котором к классическим девяти частям речи добавлены дополнительные (искусственные) категории. Большинство из таких категорий были добавлены для более детального описания свойств частей речи (например: степени сравнения прилагательных, виды местоимений и т.д.) и сохранения смысловой нагрузки. В других случаях добавление категорий вызвано необходимостью выявления и правильной разметки сложных конструкций, например: 1) *барғым келеді* – хочу пойти, 2) *ертең келеді* – приедет завтра. В первом случае *келеді* является составной частью сложной конструкции, а во втором – самостоятельным глаголом.

Разработанная разметка является позиционной, согласно которой лексическая метка состоит из основной метки (развернутая часть речи) и закодированной строки грамматических свойств. Например, слово *алмаларым* (мои яблоки) может быть размечено следующим образом: *алмаларым* / ZEP_A0N1S1P7C1: A0 1 – неодуш.; N1 – мн. ч.; S1 – принадлежность первому лицу в ед. ч.; P7 – третье лицо, мн. ч.; C1 – именительный падеж.

4. Методика работы с аннотаторами-разметчиками

Процесс создания наборов меток неразрывно связан с непосредственной разметкой текста. При разметке корпуса казахского языка [2] было задействовано девять аннотаторов и двое проверяющих. Проверка велась наряду с разметкой и сопровождалась регулярными сессиями работ с ошибками с целью синхронизации разметки. Всего было проверено около 10% всех размеченных данных и около 6% слов было исправлено, т.е. процент ошибок был сравнительно небольшим. Также для синхронизации разметки мы снабдили интерфейс разметки рекомендательной системой (РС), основанной на рекомендации вариантов разметки для ранее размеченных словоформ. Таблица 2 содержит характеристики процесса разметки без использования и с использованием РС.

Таблица 2.

Различные характеристики процесса аннотации

	без РС	с РС
Коэффициент согласия между аннотаторами	0,81	0,84
Абсолютная ошибка	0,08	0,07
Средняя скорость, слов/час	212,1	322,6

5. Заключение

В настоящей статье кратко описана методика создания синтаксического, лексического (части речи) и морфологического наборов разметок для казахского языка. Основное значение уделено лексической разметке и целесообразности интегрирования различных грамматических свойств в данный вид разметки. Первоначальные эксперименты на сравнительно небольшом наборе данных показывают, что наибольшая точность лексического разбора достигается при использовании разметки с минимальным набором грамматических свойств.

* * * * * * * * * *

1. *Макажанов А., Махамбетов О., Сабырғалиев И., Есенбаев Ж.* Разработка синтаксического, лексического и морфологического наборов разметок для казахского языка в сетях // Материалы конференции TEL-2014. – Казань, 2014. – С. 124 – 130.
2. *Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A.* Assembling the Kazakh language corpus // In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. – P. 1022 – 1031.
3. *Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Бурибаева А.К., Карабалаева М.Х.* Интеллектуальный морфологический анализатор, основанный на семантических сетях // Материалы международной научно-технической конференции OSTIS-2012. – Минск: Минск БГУИР, 2012. – С. 397 – 400.
4. *Hajič J., Hladkʹa B.* Tagging inflective languages: prediction of morphological categories for a rich, structured tagset // In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, 1998. – Volume 1. – P. 483 – 490.
5. *Hana J., Feldman A.* A positional tagset for Russian // In Proceedings of the 7th International Conference

on Language Resources and Evaluation. – Malta, 2010. – P. 1277 – 1284.
6. *Marcus M.P., Marcinkiewicz M.A., Santorini B.* Building a large annotated corpus of English: the Penn treebank. // Computational Linguistics, 1993. – Volume 2. – P. 313 – 330.
7. *Oflazer K., Say B., Hakkani-T¨ur D.Z., T¨ur G.* Building a turkish treebank // In Treebanks, Springer, 2003. – P. 261– 277.
8. *Slav Petrov S., Das D., McDonald R.* A universal part-of-speech tagset // In Proceedings of the 8th International Conference on Language Resources and Evaluation. – Istanbul, 2012. – P. 2089 – 2096.

DESIGNING SYNTACTIC, LEXICAL, AND MORPHOLOGICAL TAGSETS FOR GRAMMATICAL LABELING OF KAZAKH TEXTS

A.O.Makazhanov, O.E.Makhambetov, I.M.Sabyrgaliyev, Z.A.Yessenbayev

In this paper we describe the process of designing syntactic, lexical (POS), and morphological tagsets for the Kazakh language. Morphological tagsets provide labels for encoding basic grammatical properties of nominals and verbs. Derivational suffices are labeled as transitions between corresponding parts of speech. We have designed a flexible positional POS tagset that, depending on the application, can be configured to contain certain amount of morphological details. Our syntactic tagset comprises a compact set of syntactic categories well-defined in a classical grammar. The present work is an extended version of a paper [1] published in the proceedings of the "Theory, Engineering, Language" (TEL 2014) conference. The original material is extended with an entirely new section on the methodology of the annotation process.

Key words: computational linguistics, grammatical tagset design, syntax, parts of speech, morphology.

* * * * * * * * * *

1. *Makazhanov A., Mahambetov O., Sabyrgaliev I., Esenbaev Zh.* Razrabotka Sintaksicheskogo, Leksicheskogo i Morfologicheskogo Naborov Razmetok dlja Kazahskogo Jazyka v setjah // Materialy konferencii TEL-2014. – Kazan', 2014. – S. 124 – 130. (In Russian)
2. *Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A.* Assembling the Kazakh language corpus // In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. – R. 1022 –1031.(In English)
3. *Sharipbaev A.A., Bekmanova G.T., Ergesh B.Zh., Buribaeva A.K., Karabalaeva M.H.* Intellektual'nyj morfologicheskij analizator, osnovannyj na semanticheskih setjah // Materialy mezhdunarodnoj nauchno-tehnicheskoj konferencii OSTIS-2012. – Minsk: Minsk BGUIR, 2012. – S. 397 – 400. (In Russian)
4. *Haji˘c J., Hladk´a B.* Tagging inflective languages: prediction of morphological categories for a rich, structured tagset // In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, 1998. – Volume 1. – R. 483 – 490. (In English)
5. *Hana J., Feldman A.* A positional tagset for Russian // In Proceedings of the 7th International Conference on Language Resources and Evaluation. – Malta, 2010. – P. 1277 – 1284. (In English)
6. *Marcus M.P., Marcinkiewicz M.A., Santorini B.* Building a large annotated corpus of English: the Penn treebank. // Computational Linguistics, 1993. – Volume 2. – R. 313 – 330. (In English)
7. *Oflazer K., Say B., Hakkani-T¨ur D.Z., T¨ur G.* Building a turkish treebank // In Treebanks, Springer, 2003. – R. 261 – 277. (In English)
8. *Slav Petrov S., Das D., McDonald R.* A universal part-of-speech tagset // In Proceedings of the 8th International Conference on Language Resources and Evaluation. – Istanbul, 2012. – P. 2089 – 2096. (In English)

* * * * * * * * * *

Макажанов Айбек Омиржанович – магистр, младший научный сотрудник, Частное Учреждение «NURIS», Назарбаев Университет.

010000, Казахстан, Астана, пр. Кабанбай батыра, 53, каб. 9321.
E-mail: aibek.makazhanov@nu.edu.kz

Makazhanov Aibek Omirzhanovich – MSc., junior researcher, Private Institution «NURIS», Nazarbayev University.

53 Qabanbay Batyr Ave, Office #9321, Astana, 010000, Kazakhstan
E-mail: aibek.makazhanov@nu.edu.kz

Махамбетов Олжас Еркенович – магистр, научный сотрудник, Частное Учреждение «NURIS», Назарбаев Университет.

010000, Казахстан, Астана, пр. Кабанбай батыра, 53, каб. 9321.
E-mail: omakhambetov@nu.edu.kz

Makhambetov Olzhas Erkenovich – MSc., researcher, Private Institution «NURIS», Nazarbayev University.

53 Qabanbay Batyr Ave, Office #9321, Astana, 010000, Kazakhstan
E-mail: omakhambetov@nu.edu.kz

Сабыргалиев Ислам Магзомович – бакалавр, стажер-исследователь, Частное Учреждение «NURIS», Назарбаев Университет.

010000, Казахстан, Астана, пр. Кабанбай батыра, 53, каб. 9321.
E-mail: islam.sabyrgaliyev@nu.edu.kz

Sabyrgaliyev Islam Magzomovich – BSc., trainee researcher, Private Institution «NURIS», Nazarbayev University.

53 Qabanbay Batyr Ave, Office #9321, Astana, 010000, Kazakhstan
E-mail: islam.sabyrgaliyev@nu.edu.kz

Есенбаев Жандос Аманбаевич – магистр, научный сотрудник, Частное Учреждение «NURIS», Назарбаев Университет.

010000, Казахстан, Астана, пр. Кабанбай батыра, 53, каб. 9321.
E-mail: zhyessenbayev@nu.edu.kz

Yessenbayev Zhandos Amanbayevich – MSc., researcher, Private Institution «NURIS», Nazarbayev University.

53 Qabanbay Batyr Ave, Office #9321, Astana, 010000, Kazakhstan
E-mail: zhyessenbayev@nu.edu.kz

Поступила в редакцию 05.03.2014