М. Копотев, Л. Пивоварова
M. Kopotev, L. Pivovarova

«НЕ ДО X»: АЛГОРИТМ ВЫЯВЛЕНИЯ УСТОЙЧИВЫХ ПАРАМЕТРОВ В СОЧЕТАНИЯХ СЛОВ[1]

NOT UP TO X: ALGORITHM FOR STABLE LEXICAL / GRAMMATICAL FEATURES EXTRACTION

Аннотация. В докладе представлен алгоритм, автоматически определяющий устойчивые лексические или морфологические признаки заданного элемента n-грамма и выстраивающий их в иерархическом порядке. Это достигается путем вычисления силы всех возможных связей между токеном и его характеристиками в искомом запросе. В результате, получается морфологический и лексический профиль определенного запроса, который включает ранжированные по шкале стабильности категории и их значения.

Abstract. This paper presents an algorithm that allows issuing a query pattern, collects multi-word expressions (MWEs) that match the pattern, and then ranks them in a uniform fashion. This is achieved by quantifying the strength of all possible relations between the tokens and their features in the MWEs. As a result, we obtain morphological and lexical profiles of a given pattern, which includes the most stable category of the pattern, and their values.

Язык как таковой можно рассматривать как «конструктикон» (constructicon[2]), в природу которого заложена спаянность единиц разного уровня и отсутствие границ между ними. По этой причине разработка формальных методов для

[1] Мы благодарим Р. Янгарбера (Хельсинки) и Н. Кочеткову (Москва), без участия которых настоящее исследование было бы невозможным. За возможность работать со «снятником» НКРЯ мы благодарим разработчиков корпуса, особенно Е. В. Рахилину и О. А. Ляшевскую.

[2] *Goldberg A. E.* Constructions at work: The nature of generalization in language. Oxford University Press, 2006.

измерения силы морфологических и лексических отношений между словами кажется нам одновременно и важной, и проблематичной. Предлагаемый подход в целом призван отвечать на вопрос, в чем причина того, что слова встретились вместе, – в их морфологических особенностях, лексической совместимости или в комбинации того и другого. Приведем примеры.

В сочетании «*не до X*» самой устойчивой морфологической категорией является падеж, для которой значение родительного падежа реализуется у большого числа словоформ, а значение второго родительного (партитива) реализуется только в нескольких конкретных лексемах. Таким образом, студент или исследователь, желающий выяснить, какие значения может принимать X в шаблоне «*не до X*», должен, в идеале, получить следующий ответ: самый устойчивый параметр для *X* – падеж, при этом:

- «не до + S.gen» – это сочетание с открытым списком лексем (коллигация);
- «не до + S.gen2» – это устойчивые сочетания типа *не до смеху*, *не до жиру* (коллокации).

Таким же образом можно сказать, что в сочетании «*как X на сковородке*» самый устойчивый морфологический признак – одушевленность существительного. В сочетании «*X знает что*» наиболее устойчивыми являются мужской род существительного и ряд конкретных лексем: *БОГ, ЧЕРТ* и т.д. и т.п.

Для автоматического решения такого рода задач мы разработали алгоритм, основанный на вычислении энтропии и дивергенции относительно нормального или выборочного распределения всех параметров для данного сочетания слов. На вход алгоритма подается произвольная n-грамма (где n=2-4), где одна или больше позиций остаются незаполненными. Алгоритм находит ответы на следующие вопросы:

- какая морфологическая категория оказывается наиболее устойчивой для этой позиции?

- какое значение этой морфологической категории наиболее устойчиво?
- и наконец, что устойчивее: конкретные токены или морфологические параметры с открытым списком лексем?

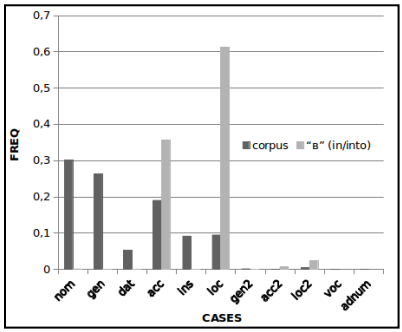
Статистические модели, использованные в данной работе, помогают распределить частоты морфологических признаков и лексических единиц определенного шаблона на единой шкале, с тем чтобы определить наиболее стабильные параметры. Например, алгоритм принимает на входе лемму «*греть*» и производит упорядочение списков ожидаемого совместного вхождения, например, устойчивое выражение «*греть душу*», словосочетание «*греть воду*» и коллигация «*греть + N.acc*». Токен или лемма не являются единственными возможными вариантами на входе алгоритма, поэтому в запросе может содержаться указание на часть речи или любая комбинации морфологических параметров – в любом случае система находит все левые или правые контексты для заданного запроса и организует их в соответствии с их морфологической и лексической устойчивостью.

Метод

Говоря кратко, разработанная модель определяет разницу в распределении параметров в целом по корпусу и в определенном шаблоне.

Например, на рис. 1 представлено распределение падежей существительного в корпусе (темно-серые плашки) и после предлога *в* (светло-серые плашки). На рис. 2 показано соответствующее распределение родов существительного. Как видно, значения рода распределены более или менее равномерно как в корпусе, так и в ограниченном контексте с предшествующим предлогом *в*. Распределение падежей, напротив, совершенно иное: винительный и предложный падежи существенно чаще встречаются после этого предлога, чем любые другие. Это связано с тем известным фактом, что предлоги

контролируют падеж управляемого существительного, но не его род. Именно в этом состоит основанная идея нашей модели.



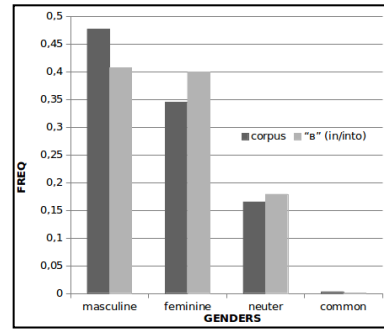


Рис.1 *Рис.2*

Для измерения этого распределения мы используем нормализованную дивергенцию Кульбака-Лейблера[3]. Параметры с наибольшим значением нормированной дивергенции считаются максимально связанными шаблоном. Для определения существенных значений выбранного параметра используется мера отклонения (weirdness measure, frequency ratio), которая

[3] *Bigi B.* Using Kullback-Leibler distance for text categorization. // Advances in Information Retrieval. 2003. P. 305–319.

была предложена в работе[4] и использована в ряде других[5]. В целом, алгоритм работает следующим образом[6]:

- поиск всех токенов, которые появляются в шаблоне запроса, и их группировка в соответствии с частеречными тегами;
- расчет внутри каждой POS-группы нормализованной дивергенция Кульбака-Лейблера для всех параметров: тегов морфологической разметки, лемм и токенов; параметры, которые показывают максимальное расхождение с общекорпусным распределением, считаются наиболее значимыми для шаблона;
- значения каждой категории сортируются в соответствии с мерой отклонения: если значение меры составляет меньше 1, то значение считается случайным.

Эксперименты

В качестве материала для исследования мы использовали данные, извлеченные из подкорпуса со снятой омонимией НКРЯ (т.н. «снятник» НКРЯ, 5 944 188 токенов).

---

[4] *Ahmad K., Gillam L., Tostevin L.* University of Surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder) //The Eighth Text REtrieval Conference (TREC-8). 1999.

[5] Например: *Chetviorkin I., Loukachevitch N.* Extraction of Domain-specific Opinion Words for Similar Domains // Information Extraction and Knowledge Acquisition. – 2011. – С. 7; *Крылова И.В., Пивоварова Л.М., Савина А.Н., Ягунова Е.В.* Исследование новостных сегментов российской «снежной революции»: вычислительный эксперимент и интуиция лингвистов // Сборник трудов конференции «Понимание в коммуникации». 2012.

[6] Более подробное описание см.: *Kopotev M., Kochetkova N., Pivovarova L. & Yangarber R.* Automatic detection of stable grammatical features in n-grams. // The 9th Workshop on Multiword Expressions, NAACL2013 (Atlanta, June 13/14, 2013). [В печати; доступно по адресу: **http://www.helsinki.fi/~kopotev/naacl-2013-mwe.pdf**]

В настоящий момент разработан только основной алгоритм, который систематически проверен на списке из 25 непроизводных предлогов, которые обладают легко предсказуемым морфосинтаксическим свойством – падежным управлением. Предсказания модели совпадают с этим ожидаемым для всех из них, другими словами наибольшее значение дивергенция Кульбака-Лейблера принимает для категории падежа, для которой «мера отклонения» во всех случаях больше единицы. Это означает 100% точности (precision) и полноты (recall) в этом случае. При проверке конкретного значения категории мы обнаружили, что алгоритм предсказывает правильные падежи для 21 из 25 предлогов[7].

Таким образом, мы можем говорить, что алгоритм достаточно надежно предсказывает коллигации, или устойчивые сочетания лемм/токенов и морфологических признаков. Однако система, над которой мы работаем, предназначена для обработки как морфосинтаксической, так и лексической совместной встречаемости, рассматривая их как единый континуум без четких границ. Следующий шаг, над которым мы начинаем работать, – это устойчивость токенов / лемм в шаблоне. Было установлено[8], что в корпусе без морфологической разметки даже простая сортировка по частоте работает достаточно хорошо. Поскольку наши данные содержат богатую морфологическую информацию, при выявлении коллокаций мы опираемся именно на нее. Опишем наш подход на конкретных примерах.

---

[7] Подробный анализ ошибок см.: *Kopotev M. et al*. Указанное сочинение (2013).

[8] *Wermter J., Hahn U.* You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006. P. 785–792.; *Kilgarriff A., Rychly P., Kovar V., Baisa V.* Finding Multiwords of More Than Two Words. // Proceedings of EURALEX. 2012.

Ничтоже сумняшеся. Существуют коллокации, в которых грамматические параметры оказываются менее устойчивыми, чем токены. К самым очевидными примерам относится, в частности, оборот *ничтоже сумняшеся*, в котором грамматических параметров нет вовсе, а после *ничтоже* возможен только один токен. Естественно, что в запросе *«ничтоже X»* максимальное значение дивергенции показывается класс Токены и его значение *сумняшеся*. Запрос на установление левого контекста *«X сумняшеся»* тоже возвращает максимальную дивергенцию для Токенов. Это, конечно, самый простой случай.

Слово в слово оказывается более сложным случаем. На месте *X* в запросе *«слово в X»* возможны единицы, в разной степени устойчивости: случайные (*слово в театр*), частотные свободные сочетания (*слово в предложении*), устойчивые обороты (*слово в защиту*), идиомы (*слово в слово*). Нормализованная дивергенция определяет в качестве победителя только грамматические параметры, при этом самыми устойчивыми оказываются категория одушевленности, падежа и рода:

Таблица 1. Нормализованная дивергенция в запросе «слово в»

Категория	**Нормализованная дивергенция**
Одушевленность	0,42
Падеж	0,39
Род	0,36
Число	0,20
Леммы	0,004

Как мы видим, категория числа и особенно леммы проигрывают другим категориям. В то же время выявить победившее значение в первых трех категориях не удается: мера отклонения не превышает единицы ни для одного из значений одушевленности, рода и падежа (Табл. 2).

Таблица 2. Максимальные меры отклонения грамматических значений в запросе «слово в»

Значение категории	**Меры отклонения**
Одушевлённость: inan	0,74
падеж: acc	0,13
род: n	0,12
Падеж: loc	0,028

Содержательно это значит, что ни одно из грамматических значений не является достаточно стабильным, чтобы выделить коллигацию вида «*слово в* [gram.tag]»[9]. В то же время значения неодушевленности и, в меньше степени, аккузатива являются наиболее устойчивыми из всего набора возможных признаков. Мы полагаем, что именно эти сведения можно использовать для определения коллокации в тексте.

Идея, лежащая в основе следующего шага алгоритма, основана на использовании тех данных, которые мы получили на предыдущем этапе. В этом мы опираемся на идею грамматических профилей, предложенную С. Грайсом и Д. Дивьяк и развитую Л. Яндой и О. Ляшевской[10]. Наша реализация этого подхода основана на том, что распределение токена/леммы в шаблоне подсчитывается с учетом грамматических признаков, получивших наибольшие значения меры отклонения, то есть являются наиболее специфичными для шаблона. Так, в указанном примере устойчивость лемм в шаблоне подсчитывается не относительно всех токенов или всех

[9] В шаблоне «*в X»,* естественно, побеждает категория падежа и значения *acc/acc2/loc/loc2*.

[10] *Gries S. Th., Divjak D.S.* Behavioral profiles: a corpus-based approach towards cognitive semantic analysis. // New directions in cognitive linguistics. 2009. P. 57–75; *Janda L., Lyashevskaya O.* Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian // Cognitive Linguistics 22:4 (2011). P. 719–763.

лемм-существительных в корпусе, а только относительно токенов, имеющих такой же грамматический профиль, например, S.inan.acc. Проверим это.

Во-первых, мера отклонения для лемм, подсчитанная относительно общекорпусной, дает ожидаемо плохой результат. Лемма *отдельность* побеждает, потому что она вообще очень редко встречается в корпусе.

Таблица 3. Отклонение от общекорпусных параметров

лемма	**Мера отклонения**
слово	151,55
адрес	232,44
отдельность	1183,06

В таблице ниже приведены данные той же меры отклонения, подсчитанной с целью эксперимента относительно подвыборок, представляющих разные грамматические профили.

Таблица 5. Мера отклонения с учетом грамматических профилей

	S.inan	**S.inan.acc**	S.sg	S.loc	S.acc
слово	111,78	**96,16**	253,26	0,00*	109,32
адрес	171,44	**76,47**	224,37	0,00*	869,33
отдельность	872,59	**0,00***	940,64	0,47	0,00*

Из этого эксперимента следует, что результат, в наибольшей мере соответствующий интуиции (*слово в слово*), получен только для профиля S.inan.acc, в который входят два самых релевантных признака, полученных на предыдущем этапе: неодушевленность и аккузатив. Во всех других случаях побеждает лемма *отдельность* – отнюдь не потому, что она характерна для этой коллокации, а потому, что она является редкой, и даже ее случайное попадание в цепочку обеспечивает ей победу. На

* В шаблоне не встретился токен с указанным грамматическим профилем.

практике использование грамматического профиля позволяет отсечь лексический материал, не поддержанный синтаксической конструкцией.

Выводы

Конечно, наша работа находится еще в самом начале. Но уже сейчас можно видеть, что извлечение коллигаций произвольной длины предложенным методом дает достаточно надежные результаты. Извлечение коллокаций с учетом грамматических профилей кажется ближайшей перспективной задачей.