Иванка Атанасова, Преслав Наков, Светлин Наков (Болгария)

СЕМАНТИЧЕСКАЯ ТЕХНИКА АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ГИПОНИМИЧЕСКИХ РЯДОВ ИЗ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЕЙ

1. Семантическая техника

Семантическая техника автоматического извлечения гипонимических рядов основывается на латентном семантическом анализе и используется, главным образом, для гипонимов, не имеющих общего терминоэлемента, хотя в списки попадают и гипонимы с общим терминоэлементом. У нее две разновидности: *семантическая техника без сегментации и семантическая техника с сегментацией.* Со своей стороны, **семантическая техника без сегментации** тоже имеет две разновидности: *семантическая техника для **толкований** терминов-гиперонимов (или терминов-гипонимов) и семантическая техника для **самих** терминов-гиперонимов (или терминов-гипонимов).*

2. Латентный семантический анализ

Латентный семантический анализ (ЛСА) – мощная статистическая техника индексации, извлечения и анализа текстовой информации, применяемая с успехом в различных областях человеческого познания за последнее десятилетие. Метод полностью автоматический и не пользуется никакими предварительно составленными словарями, семантическими сетями, базами знания, концептуальными иерархиями, грамматическими, морфологическими или синтаксическими анализаторами и др. В его основе лежит гипотеза о том, что между отдельными словами и обобщенным контекстом (предложениями, абзацами и целыми текстами), в которых они встречаются, существуют неявные (латентные) взаимосвязи, обуславливающие совокупность взаимных ограничений. Их открытие и правильное рассмотрение дают возможность ЛСА справиться успешно с синонимией и частично с полисемией: с обеими наиболее сложными проблемами при статистической обработке текстовой информации.

ЛСА двухступенчатый процесс, включающий обучение и анализ результатов. Фаза обучения начинается с формирования частотной матрицы встреч слов/терминов в документах. Следует логарифмическая трансформация, деление на энтропию по рядам, декомпозиция по сингулярной стоимости и выключение шума, в результате чего получается небольшое число значимых факторов, чаще всего между 50 и 400. Таким образом, каждому документу и каждому слову сопоставляется вектор небольшой размерности одного и того же семантического пространства. Степень близости между двумя документами/словами определяется как функция соответствующих им векторов, чаще всего посредством косинуса угла, заключенного между ними [Berry, 1993; Deerwester, Dumais, Furnas, Laundauer, Harshman, 1990, pp. 391-447; Laudauer, Foltz, Laham, 1998, pp. 259-284; LSA, 1990-2001; Nakov P.–1,

2000, pp. 156-166; Nakov P.–2, 2000, pp. 189-198; Nakov P.–3, 2000; Nakov P., 2001].

Исследование проводилось на материале составленных нами *компьютерных словарей терминов изобразительного искусства (КСТИИ),* в которых каждая словарная статья состоит из слова, т. е. *однословного термина (ОТ)* или *терминологического словосочетания (ТС),* расположенного с левой стороны, и краткого толкования соответствующего термина, включающего все его значения, расположенного с правой стороны. Из них извлечены 116 русских и 118 болгарских гипонимических рядов с минимальным количеством два гипонима и максимальным количеством свыше 190 гипонимов. Гипонимы, не имеющие общего терминоэлемента, мы отыскивали с помощью ЛСА, причем болгарский и русский КСТИИ были исследованы в отдельности. В процессе построения исходной матрицы были исключены стоп-слова, т. е. слова, встречающиеся чаще всего, но лишенные собственного смысла или слишком многозначные. В эту категорию входят, главным образом, статичные элементы языка: союзы, предлоги, междометия, частицы, местоимения, количественные числительные, формы глагола р. “быть” (б. “съм”) и др. Мы исключили также и слова, встречающиеся только в одном документе (только из толкований), так как они не могут способствовать оценке близости. Таким образом, число различных словоформ уменьшилось более чем в два раза – соответственно до 4369 для болгарского и 4485 для русского словарей. Наши исследования [Атанасова, Наков-1, 2001, с. 327-334; Атанасова, Наков-2, 2001; Атанасова, Наков-4, 2001] показывают, что при сильно флективных славянских языках для эффективности результатов применения ЛСА имеет значение правильное определение и отождествление форм одного и того же слова (сегментация). Мы сделали автоматическую сегментацию, причем число слов уменьшилось соответственно до 2263 и 2299. В процессе поисков гипонимов были использованы параллельно 4 различных семантических пространства, все с размерностью 100: по два для болгарского и русского словарей (с и без сегментации).

3. Семантическая техника без сегментации

Модулю ЛСА подаются *толкования* гиперонима или одного из гипонимов данного гипонимического ряда, т. е. информация правой стороны соответствующего терминологического словаря, и компьютер составляет список, включая весь словарь. По своему усмотрению, исследователь выделяет отрезок, обычно содержащий большое количество терминов, не являющихся согипонимами, например: *р. дерево –... сосна, липа, грецкий орех (в. орех) *гниль, *червоточина, кипарис…; бирмит (гипер. янтарь) – руменит, *дымчатый кварц…; б. дърво - ...бряст, трепетлика, слива, *чакъл, чимшир, *сахтиян…; пушкарство (гипер. оръжейничество) – саблярство (д. чалъклийство), *пръстенджийство, ножарство (д. бучакчийство), *карнавал, *килимарство….*

В целях извлечения максимального количества гипонимов, компьютерной программе подаются последовательно толкования гиперонима и нескольких гипонимов, причем согипонимы в списке результатов меняют свои места. Результаты сравниваются, и гипонимический ряд дополняется.

При ЛСА в рассматриваемые компьютерные списки попадают и согипонимы с общим терминоэлементом, так как при этой технике не учитывается формальное выражение ОТ и ТС, например: ***р. агат*** *- … белый* ***агат*** *(д. кахолонг), желтый* ***агат****, оникс, сардоникс…;* ***б. халцедон*** *- …обикновен* ***халцедон****, карнеол, хелиотроп, хелиопраз…*

4. Сегментация

Современная статистическая обработка текстовой информации основывается исключительно на анализе частоты встречи отдельных слов (иногда словосочетаний) как основного носителя языка в анализируемом наборе текстов. В этом смысле основной проблемой является дефиниция – что представляет собой слово с точки зрения применяемого алгоритма. При большинстве алгоритмов наблюдается улучшение результатов при отождествлении форм одного и того же слова.

Основная цель сегментации – привести различные формы слов к какой-нибудь основе (префикс + корень), чаще всего путем устранения суффиксов и окончаний. Алгоритмы для сегментации почти никогда не устраняют префиксов, потому что это легко может привести к коренному изменению значения слова (напр. *б. живописване* и *надживописване*). Многообразие словоформ обязано, с одной стороны, флективной, а, с другой – вариативной морфологии, причем современные алгоритмы для сегментации атакуют обе. Так, например, к классу болгарского слова *б. портрет* относятся еще флективные формы *портрета, портретът, портрети, портретите* и др., а русского слова *р. портрет – портрета, портрету, портретом, портрете, портреты, портретов, портретам, портретами, портретах.* В лингвистике совокупность всех грамматических форм данного слова, соотносящихся с одним из ее значений, называются лексико-семантическими вариантами (ЛСВ) [Новиков, 1982, с. 113] или лексико-грамматическими аллолексами [Вътов, 1998, с. 15]. Большинство алгоритмов для сегментации включают сюда и некоторые вариативные формы с различным лексическим значением, например *р. портретист, портретистка, портретировать* ‘создавать портрет данного лица’, *портретируемый, портретирование, портретная* ‘галерея с портретами’; *б. портретист, портретистка, портретирам, портретиране, портретиран, портретна,* а также и параллельные вариативные формы с тождественным лексическим значением, например *б. портретирам – портретувам, портретиран – портретуван, портретиране – портретуване.* В лингвистике вариативные формы, имеющие различное значение, рассматриваются как отдельные, однокоренные слова, а на лексическом уровне – как отдельные лексемы [Новиков, 1982, с. 114; Вътов, 1998, с. 15]. Как известно, в терминологии параллельные вариативные формы, т. е. аллолексы (фонетические, ор-

фографические, морфологические и др.) с тождественным лексическим значением называются вариантами. Практически алгоритмы для сегментации включают термины с общим терминоэлементом, выраженным корнем или основой.

Основные исследования в области сегментации проводились для английского языка. Разнообразие алгоритмов большое: самые простые сводятся к простому устранению множественного числа "-s" (и эвентуально форм глаголов, оканчивающихся на "-ed" и "-ing"), а более сложные основываются на наборе правил. Классические алгоритмы Ловинса [Lovins, 1968, pp. 22-31] и Портера [Porter, 1980, pp. 130-137] включают соответственно 260 и 60 правил. Современные алгоритмы проводят сегментацию на основе словаря или эксплуатируют статистическую информацию, базирующуюся на наблюдениях над текстом, чаще всего в комбинации с алгоритмами, основывающимися на морфологических правилах.

Несмотря на то, что результаты ее применения противоречивы (см. [Harman, 1991, с. 7-15]), последние исследования показывают, что сегментация приводит к улучшению, хотя и далеко не во всех случаях настолько заметному. Результаты варьируют значительно у различных авторов: от 1-3% [Hull, 1996, pp. 70-84] до 30-40% [Krovetz, 1993, pp.191-202]. Эксперименты показывают, что сегментация полезна не только для английского, но и для других языков, например словенский [Popovic, 1992, pp. 384-390] и голландский [Kraaij, 1996, pp. 40-48].

5. Семантическая техника с сегментацией

В результате целого ряда экспериментов и анализов [Атанасова И., Наков П., 2001-1, 2001-2, 2001-4] мы установили, что если подавать последовательно модулю ЛСА толкования гиперонима и нескольких гипонимов данного гипонимического ряда без сегментации и с сегментацией, можно получить разнообразные результаты. Мы использовали стандартные мерки *пресижон* (precision) и *рикол* (recall) для оценки результатов.

Наши исследования показывают более высокие усредненные результаты открываемости гипонимов в гипонимических рядах в терминологии ИИ с помощью ЛСА *без сегментации* как для рикола (р. 66,05%; б. 69.04%), так и для пресижона (р. 46,72%; б. 50,18%). По нашим наблюдениям, в отличие от английского языка в славянских языках (болгарском и русском) *при сегментации* усредненные результаты более низкие как для рикола (р. 56,77%; б. 64,09%), так и для пресижона (р. 36,82%; б. 43,03%).

6. Семантическая техника для терминов-гиперонимов (или терминов-гипонимов)

Дуалистический характер ЛСА позволяет измерять близость не только между *толкованиями* (текстами), но и между *отдельными понятиями* (заданными словом или словосочетанием, т. е. ОТ или ТС). Проведены эксперименты с целью сравнить эффективность обеих семантических техник – семанти-

ческой техники для *толкований* терминов-гиперонимов (или терминов-гипонимов) и семантической техники для *самих* терминов-гиперонимов (или терминов-гипонимов).

Наши исследования показывают более высокие усредненные результаты извлечения гипонимических рядов из КСТИИ при ЛСА толкований гипонимов или гиперонимов для русской терминологии ИИ (пресижон: 48,91%; рикол: 54,29%) и более низкие результаты для болгарской терминологии ИИ (пресижон: 48,81%; рикол: 54,16%). При ЛСА терминов (гипонимов или гиперонимов) получаются более высокие усредненные результаты для гипонимических рядов болгарской терминологии ИИ (пресижон: 50,92%; рикол: 55,98%) и более низкие результаты для гипонимических рядов русской терминологии ИИ (пресижон: 48,69%; рикол: 54,36%).

Эксперименты показывают, что ЛСА довольно хорошо справляется сам с задачей построения хороших векторов терминов, и эксплицитная подача контекста в виде толкования скорее мешает ему, чем помогает.

7. Выводы

Настоящее исследование приводит нас к следующим выводам:

1. Семантическая техника автоматического извлечения гипонимических рядов экономит много времени и сил, помогает исследователю-лингвисту быстрее получить более точные результаты. Эффективность семантической техники в большей степени зависит от умения лингвиста удачно подобрать и подать нужный языковой материал для анализа и правильно растолковать полученные результаты.

2. Как правило, результаты, полученные при применении семантической техники автоматического извлечения гипонимических рядов, нуждаются в дополнительном анализе посредством других методов научного исследования. Только лингвист может убрать ненужный материал, дополнить его, проанализировать, сделать соответствующие выводы. Предложенная семантическая техника, как и другие специальные компьютерные технологии, очень нужные и надежные помощники, которые ни в коем случае не уменьшают роль лингвиста в научном исследовании.

3. Семантическая техника может быть использована для сходных лингвистических исследований как конкретных, так и сопоставительных в области лексикологии и семантики.

Литература

Атанасова И., Наков П.-1. Автоматично извличане на хипоними от терминологични речници. – ВВМУ “Н.И. Вапцаров”. Морски научен форум. Приложна лингвистика и чуждоезиково обучение, т. 3. Варна, 2001.

Атанасова И., Наков П.-2. Ролята на сегментацията при автоматично извличане на хипоними от терминологични речници. – ВТУ “Св. Св. Кирил и Методий”. Научна сесия “Съвременни постижения на филологическите науки и университетското обучение по чужд език”. В. Търново, 2001.

Атанасова И., Наков П.-3. Термин и документ от гледна точка на латентния семантичен анализ. – ВВОВУ "Васил Левски". Научна конференция '2001' "Технологии, сигурност и екология". Научни трудове, кн. № 69. В. Търново, 2001.

Атанасова И., Наков П.-4. Факторы, влияющие на автоматическое извлечение гипонимов из терминологических словарей с помощью латентного семантического анализа. – ШУ "Епископ Константин Преславски". Юбилейна научна конференция. Шумен, 2001.

Вътов В. Лексикология на българския език. Лексемика. Ономастика. Фразеология. Лексикография. Велико Търново, 1998.

Berry M., Do T., O'Brien G., Krishna V., and Sowmini Varadhan. SVDPACKC (Version 1.0) User's Guide, 1993.

Deerwester S., Dumais S., Furnas G., Laundauer T., Harshman R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Sciences, 41, 1990.

Harman D. How effective is suffixing? In Journal of The American Society of Information Science. Vol. 42, No 1, 1991.

Hull, D. Stemming Algorithms: A Case study for detailed evaluation. In Journal of The American Society of Information Science. Vol. 47, No 1, 1996.

Kraaij W. Viewing stemming as recall enhancement. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. New York, 1996.

Krovetz R. Viewing Morphology as an Inference Process. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. New York, 1993.

Laudauer T., Foltz P., Laham D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 1998.

Lovins J. Development of a stemming algorithm. Mech. Trans. And Comp. Ling. 11., 1968.

LSA. 1990-2001, see http://lsa.colorado.edu

Nakov P.–1. Getting Better Results with Latent Semantic Indexing. In Proceedings of the Students Presentations at ESSLLI-2000. Birmingham, UK, 2000.

Nakov P.–2. Web-personalisation using extended Boolean operations with Latent Semantic Indexing. In Proc. of AIMSA-2000 (Artificial Intelligence: Methodology, Systems and Applications). Lecture Notes in Artificial Intelligence 1904, Springer. Varna, Bulgaria, 2000.

Nakov P.–3. Latent Semantic Analysis of Textual Data. In Proceedings of CompSysTech'2000. Sofia, Bulgaria, 2000.

Nakov P. Latent Semantic Analysis for Bulgarian literature. In Proceedings of Spring Conference of Bulgarian Mathematicians Union. Borovetz, 2001.

Popovic M., Willett P. The Effectiveness of Stemming for Natural Language access to Slovene Textual Data. In Journal of The American Society of Information Science. Vol. 43, No 5, 1992.

Porter M. An algorithm for suffix stripping. Program 14, 3, 1980.