РАЗРЕШЕНИЕ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ НА ОСНОВЕ ТЕЗАУРУСА ПРЕДМЕТНОЙ ОБЛАСТИ[1]

LEXICAL DISAMBIGUATION BASED ON DOMAIN-SPECIFIC THESAURUS

Лукашевич Н.В. *(louk@mail.cir.ru),*
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова
Добров Б.В. *(dobroff@mail.cir.ru), АНО Центр информационных исследований*

В статье описывается существующая автоматическая процедура выбора значений на основе структуры Общественно-политического тезауруса, а также эксперимент по оценке качества разрешения лексической многозначности для описанного алгоритма.

1. Введение

Одной из серьезных проблем автоматической обработки текстов является проблема разрешения лексической многозначности [5, 1]. С 1998 года для тестирования систем автоматического разрешения лексической многозначности проводится специальная конференция SENSEVAL (www.senseval.org).

Исследования методов разрешения лексической многозначности как отдельной задачи обычно делятся на два направления: разрешение лексической многозначности некоторой совокупности слов (чаще всего, несколько десятков) и разрешение лексической многозначности всех слов текста [7].

Между тем необходимо обозначить еще одно направление исследований в этой области, связанное с количественными характеристиками рабочего множества многозначных слов. Это направление заключается в необходимости качественного разрешения лексической многозначности для нескольких сотен - нескольких тысяч слов и связано с задачей использования в автоматической обработке текстов онтологий или тезаурусов в некоторой конкретной предметной области.

Действительно, несмотря на то, что предполагается, что идеальный термин должен быть однозначным, в реальных текстах можно встретить значительное количество случаев неоднозначного употребления терминов. Многозначность терминов связана с тем, что термины подчиняются законам функционирования языка и на них действуют многие законы, относящиеся к общезначимой лексике языка, например, возникает явление регулярной многозначности, например, *смазка* (*процесс*) – *смазка* (*вещество*) [3].

Кроме того, некоторые термины могут совпадать со словами литературного языка, употребляться в текстах как в терминологическом, так и нетерминологическом значении, и, следовательно, возникает необходимость определения, в каком из значений употреблен термин в том или ином фрагменте текста.

Различные алгоритмы разрешения лексической многозначности на основе тезаурусной структуры предлагались и тестировались для тезауруса английского языка WordNet. Все известные авторам алгоритмы основываются на анализе относительного расположения соответствующих значениям данного многозначного слова синсетов WordNet и синсетов слов из текстового контекста. Для оценки близости синсетов данного многозначного слова и синсетов контекста предлагалось использовать такие параметры как:

- длина самого короткого пути по отношениям WordNet между каждым синсетом многозначного слова и синсетами контекста [14];
- глубина по иерархии WordNet [4];
- наименьший общий вышестоящий синсет [15, 12, 6];
- отнесенность к так называемым доменам – тематическим областям, приписанным синсетам WordNet'а [8, 16].

В работе [11]описывается тестирование ряда предложенных на базе WordNet методов разрешения лексической многозначности на материалах конференции SENSEVAL-2. Методы были применены для разрешения многозначности 1723 существительных коллекции. Лучший результат (точность разрешения многозначности - 39%) был получен для метрики, предложенной в работе [6].

[1] Работа частично выполнена при поддержке гранта РФФИ № 05-07-90391-в и при поддержке ООО "Яндекс" (www.yandex.ru).

Для сравнении приведем точность работы лучших систем разрешения лексической многозначности в соревновании SENSEVAL-3. Для английского языка в задаче разрешения многозначности для всех слов текста точность лучшей системы составляет 65.2% [13], для задачи определения многозначности для заданного набора слов (40 слов) – 72% [9]. Организаторы конференции подчеркивают, что лучшие результаты достигаются теми системами, которые используют комбинации нескольких классификаторов и особенно подходы, основанные на обучении по размеченным корпусам (так называемые подходы supervised disambigation).

В статье рассматриваются способы описания многозначных терминов в Общественно-политическом тезаурусе, который является подтезаурусом Тезауруса русского языка РуТез [2] и относится к широкой предметной области современной общественной жизни. В настоящее время Общественно-политический тезаурус содержит 33 тысячи понятий, 87 тысяч терминов.

Выделение Общественно-политического тезауруса в рамках большего ресурса может быть сопоставлено с подходом, возникшим при разметке WordNet предметными областями [8], когда дополнительно к набору тематических областей была введена специальная область Factotum. К этой области относятся синсеты WordNet, не входящие в конкретные тематические области. Именно область Factotum содержит наиболее трудно различимые значения и имеет больший процент многозначных слов [16]. Таким образом, Общественно-политический тезаурус в рамках тезауруса РуТез приблизительно соответствует объединению синсетов всех тематических областей WordNet без включения области Factotum.

В статье мы опишем существующую автоматическую процедуру выбора значений на основе структуры Общественно-политического тезауруса, а также эксперимент по оценке качества разрешения лексической многозначности для описанного алгоритма.

2. Представление разных значений слов в Общественно-политическом тезаурусе

В Общественно-политическом тезаурусе (далее Тезаурус) существуют два основных способа представления значений многозначных терминов.

Первым способом представления многозначности является задание одного и того же текстового входа разных понятиям тезауруса (М-многозначность). Так, например, текстовый вход *пилот* сопоставлен двум разным понятиям понятию *ЛЕТЧИК* и понятию *АВТОГОНЩИК*.

Такое представление используется для задания разных видов лексической многозначности:

- омонимии: слово *брак* соответствует таким дескрипторам как *СУПРУЖЕСТВО* и *ПРОИЗВОДСТВЕННЫЙ БРАК,*
- терминов из разных предметных областей: слово *прокат* соответствует таким понятиям как *ПРОКАТНОЕ ПРОИЗВОДСТВО* (металлургия), *КИНОПРОКАТ (кинематография), ПРОКАТ ИМУЩЕСТВА* (аренда).
- метонимии: слово *балет* относится к таким дескрипторам как *БАЛЕТНОЕ ИСКУССТВО (развитие балета), БАЛЕТНЫЙ СПЕКТАКЛЬ (смотреть балет), БАЛЕТНАЯ ТРУППА (приезд балета*). Отметим, что обычно понятия, которым соответствует один и тот же текстовый вход, образованный на основе явления метонимии, связаны между собой тезаурусными отношениями.
- метафоры: слово *сотовый* соответствует понятиям *СОТОВАЯ СВЯЗЬ* и *ПЧЕЛИНЫЕ СОТЫ.*

Второй способ представления многозначности используется в тех случаях, когда слово представлено в тезаурусе в одном значении, но поскольку известно, что оно может употребляться и в других значениях в текстах предметной области, то ему ставится специальная пометка многозначности (А-многозначность), например, такой пометкой отмечено слово *дар,* которое представлено в тезаурусе как текстовый вход к понятию *ПОДАРОК,* слово *история*, представленное как текстовый вход к понятию *ИСТОРИЧЕСКИЕ НАУКИ.*

Пометка многозначности используется для отметки географических названий, которые могут совпадать с фамилиями и именами людей, сокращениями и др., например, *Львов* (город), *Владимир* (город), *Павлово* (город в Нижегородской области).

В настоящее время в составе Общественно-политического тезауруса насчитывается 6463 многозначных терминов. Для 2204 терминов представлено 2 и более значений, многозначность остальных терминов отмечена пометкой.

3. Построение тезаурусных окрестностей для значений многозначных слов

Существующий алгоритм разрешения лексической многозначности основывается на выделении для каждого понятия так называемой «тезаурусной окрестности» этого понятия – то есть совокупности понятий тезау-

руса, которые считаются близкими по смыслу к исходному понятию, и, следовательно, могут «проголосовать» в пользу этого понятия в ситуации выбора значений многозначного термина. Близкие подход к разрешению многозначности на базе WordNet используется в работе [15].

Тезаурусная окрестность строится на основе свойств описываемых в тезаурусе отношений.

3.1. Набор отношений в Тезаурусе

В Тезаурусе имеется четыре основных типа отношения.

Первый тип отношений – таксономическое отношение НИЖЕ-ВЫШЕ, обладает свойством транзитивности и наследования.

Второе тип отношений – отношение ЧАСТЬ-ЦЕЛОЕ. Используется не только для описания физических частей, но и для других внутренних сущностей понятия таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен все свое существование являться частью для понятия-целого, а не относиться к чему-либо другому.

При таком условии удается ввести свойство транзитивности такого отношения ЧАСТЬ-ЦЕЛОЕ, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого несимметричной ассоциацией АСЦ2-АСЦ1, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, и когда одно из понятий не существует без существования другого. Например, понятие *САММИТ* требует существования понятия *ГЛАВА ГОСУДАРСТВА*.

Последний тип отношений – симметричная ассоциация связывает, например, понятия очень близкие по смыслу, но которые не были соединены в одно понятие.

3.2. Алгебра отношений в Тезаурусе

Отношения таксономии, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии. Мы называем такую совокупность понятий *«деревом-вниз»* для исходного понятия. Соответственно, совокупность понятий, вышестоящих по иерархии для данного понятия, называется его *«деревом-вверх»*.

Например, для построения «дерева вверх» действуют такие основные правила:

ВЫШЕ (X,Y)	+	ВЫШЕ (Y,Z)	=>	ВЫШЕ (X,Z)
ВЫШЕ (X,Y)	+	ЦЕЛОЕ (Y,Z)	=>	ЦЕЛОЕ (X,Z)
ВЫШЕ (X,Y)	+	АСЦ1 (Y,Z)	=>	АСЦ1 (X,Z)
ВЫШЕ (X,Y)	+	АСЦ (Y,Z)	=>	АСЦ (X,Z)
ЦЕЛОЕ (X,Y)	+	АСЦ1 (Y,Z)	=>	АСЦ1 (Y,Z)
ЦЕЛОЕ (X,Y)	+	АСЦ (Y,Z)	=>	АСЦ (X,Z)

Таким образом, понятие *C1* находится в дереве-вверх исходного понятия *C0*, если путь тезаурусных отношений от *C0* до *C1* может быть преобразован с помощью вышеописанных правил в одно отношение.

Подобные же правила записываются и для дерева-вниз.

Тезаурусная окрестность понятия *C0* определяется как объединение понятий, входящих в дерево-вверх и дерево-вниз понятия *C0*.

4. Существующий алгоритм разрешения многозначности

При автоматической обработке текста на основе Общественно-политического тезауруса первым этапом является сопоставление текста с единицами тезауруса и создание концептуального индекса, в котором указываются те понятия, которые встречались в тексте. Многозначность в этом индексе проявляется либо в сопоставлении одной и той же языковой единице разных понятий, либо в специальной пометке понятия, означающей, что текстовая единица, по которой было проведено сопоставление, является многозначной.

На втором этапе строится так называемая проекция тезауруса для анализируемого текста. Проекция включает в себя понятия индекса и тезаурусные отношения между такими понятиями, которые входят в тезаурусную окрестность друг друга.

В эту проекцию включаются и все варианты понятия, соответствующие многозначным терминам. Для них также выявляются все понятия, упомянутые в тексте и входящие в их тезаурусные окрестности. Таким образом, для каждого варианта собираются те понятия текста, которые “поддерживают” этот вариант. “Поддержка” текста проявляется двумя способами:

- в тексте встречается однозначный вариант помеченного понятия, например, упоминание в тексте словосочетания *расследование преступлений* поддерживает именно это значение у многозначного слова *следствие*.
- в тексте встречается понятие из тезаурусной окрестности неоднозначного термина, например, упоминается понятие *ОБЩЕСТВЕННАЯ ДЕЯТЕЛЬНОСТЬ* из тезаурусной окрестности неоднозначного термина *партия*.

Далее собственно и производится выбор варианта понятия для многозначного термина:

- Неоднозначность задана с помощью пометы. Если текст “поддерживает” описанное в тезаурусе значение неоднозначного термина, то соответствующее понятие включается в понятийный индекс как однозначный. В противном случае, неоднозначный термин исключается из понятийного индекса.
- Неоднозначность проявляется в соответствии одного текстового выражения нескольким понятиям. Сначала проверяется, какие из вариантов термина поддерживаются понятиями всего текста, и оставляются только “поддержанные” варианты. Если ни один из вариантов не поддерживается, то все они удаляются из понятийного индекса.

После удаления “не поддержанных” вариантов может остаться только один вариант, и, таким образом, неоднозначность разрешена.

Если же поддержано более одного варианта, то производится выбор значения именно для конкретного вхождения неоднозначного термина: выбирается тот вариант, для которого “поддерживающее” понятие находится ближе всего по тексту. Расстояние измеряется в количестве выявленных понятий между текущим вхождением неоднозначного термина и “поддерживающим” понятием.

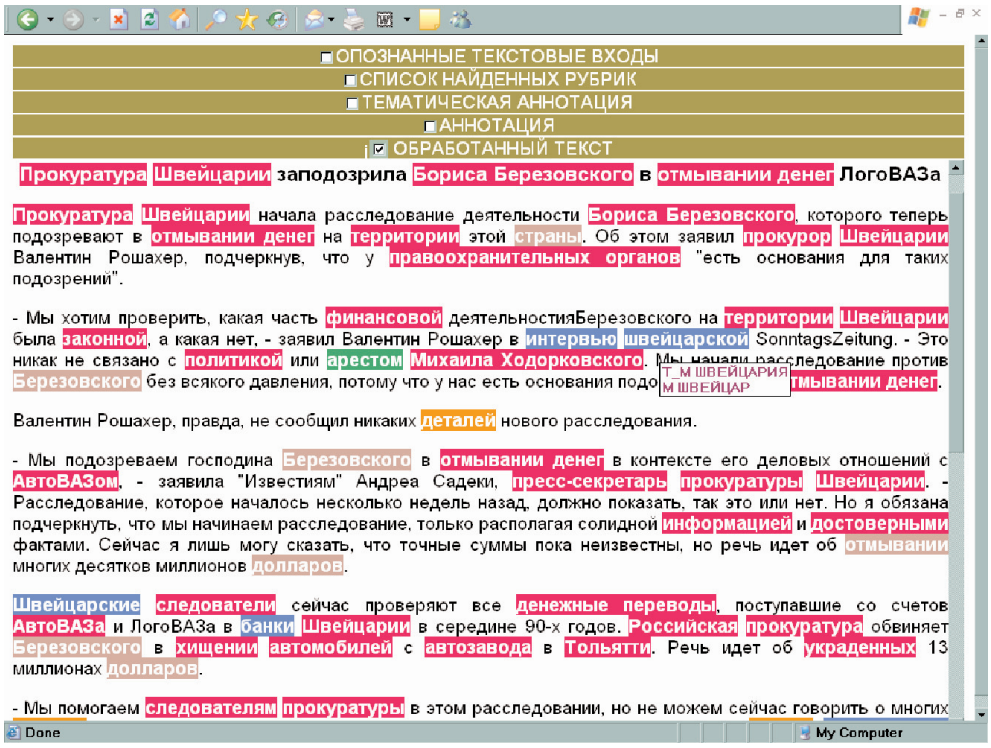

Рис.1. Фрагмент текста газетной статьи с выделением терминов из Общественно-политического тезауруса

Результаты автоматической процедуры разрешения многозначности могут быть выведены в текстовый файл специального формата. Так, например, для текста примера (Рис.1) фрагмент файла сопоставления с тезаурусом будет следующим (Таблица 1):

T_M104936;130508.921 931	ШВЕЙЦАРИЯ	ШВЕЙЦАРСКИЙ	_2
M115622;130508.921 931	ШВЕЙЦАР	ШВЕЙЦАРСКИЙ	_2
T_A6646;189163.976 984	ПОЛИТИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ	ПОЛИТИКА	_3
M6309;100526.990 996	АРЕСТ ИМУЩЕСТВА	АРЕСТ	_3
M142663;100526.990 996	ВЗЯТИЕ ПОД СТРАЖУ	АРЕСТ	_3
T130485;180997.998 1018	ХОДОРКОВСКИЙ МИХАИЛ БОРИСОВИЧ	МИХАИЛ ХОДОРКОВСКИЙ	_3
T_A122053;165720.1053 1064	БЕРЕЗОВСКИЙ БОРИС АБРАМОВИЧ	БЕРЕЗОВСКИЙ	_4
T133472;187832.1139 1153	ОТМЫВАНИЕ ДЕНЕГ	ОТМЫВАНИЕ ДЕНЬГИ	_4
A107496;136793.1234 1240	ДЕТАЛЬ (ЧАСТЬ МЕХАНИЗМА)	ДЕТАЛЬ	_5
T_A122053;165720.1325 1336	БЕРЕЗОВСКИЙ БОРИС АБРАМОВИЧ	БЕРЕЗОВСКИЙ	_6

Таблица 1. Фрагмент файла, отражающего сопоставление текста с тезаурусом

Самый правый столбец слов показывает на словарные формы слов из текста, с которыми сопоставились термины тезауруса. Левый столбец слов – это названия понятий тезауруса, которые могут соответствовать этим текстовым фрагментам.

Пометы в начале строки показывают описанный в Тезаурусе статус значения текстового фрагмента :

- T – однозначный текстовый вход,
- M – многозначный текстовый вход, для которого в тезаурусе описано более одного значения,
- A – многозначный текстовый вход, который в тезаурусе имеет только одно значение,
- T_M – выбрано одно из представленных значений многозначной языковой единицы,
- T_A – подтверждено описанное в Тезаурусе значение.

Таким образом, приведенный фрагмент файла сопоставления показывает, что слову *швейцарский* в Тезаурусе соответствует два понятия *ШВЕЙЦАРИЯ и ШВЕЙЦАР,* при этом система автоматически выбрала понятие *ШВЕЙЦАРИЯ.* Слову *арест* соответствует два понятия: *АРЕСТ ИМУЩЕСТВА и ВЗЯТИЕ ПОД СТРАЖУ* – система не смогла выбрать ни одно из значений. Система также не подтвердила значение встретившегося в тексте слова *деталь*, как части механизма (анализировался следующий фрагмент: «*не сообщил никаких деталей нового расследования*») и подтвердила значение слова *политика* как *ПОЛИТИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ.*

5. Тестирование алгоритма разрешения многозначности по Общественно-политическому тезаурусу

Тестирование алгоритма разрешения многозначности проводилось на материалах газет. Предварительно, случайным образом было выбрано несколько дат. Из коллекции Университетской информационной системы РОССИЯ (www.cir.ru) были выгружены газетные публикации, относящиеся к выбранным датам. Набор газетных публикаций включает полные номера газет «Известия», «Ведомости», «Независимая газета», «Комсомольская правда». Каждый номер содержит несколько десятков статей. Средний размер статьи около 5 Кб. На рис.1 показан фрагмент одной из публикаций. Слова и словосочетания, которые были сопоставлены с терминами Общественно-политического тезауруса, выделены, таким образом можно видеть среднее покрытие газетного текста терминами Общественно-политического тезауруса.

Для определения качества разрешения лексической многозначности необходимо было выполнить эталонную разметку найденных терминов по значениям. Для каждого документа были получены файлы, подобные фрагменту файла, приведенному в разделе 4. Такие файлы были исправлены экспертами, которые пометками указали, какое значение нужно было выбрать в каждом конкретном случае.

После получения эталонных файлов они были автоматически сопоставлены с результатами работы программы разрешения многозначности. Были выделены следующие случаи соответствия (несоответствия) эталонной разметки и результирующего файла работы программы:

1) Значение было выбрано правильно;
2) Значение не было выбрано, и это было правильно;
3) Значение было выбрано неправильно;
4) Значение не было выбрано, и это было неправильно;
5) Система выбрала один из правильных вариантов.

В качестве правильных решений системы рассматривались виды соответствия 1), 2) и 5).

В процессе эксперимента вручную было размечено 197 документов, что соответствует полным номерам газет «Известия», «Независимая газета», «Ведомости», «Комсомольская правда» от 19 ноября 2003 года. Результаты работы алгоритма разрешения многозначности по каждому из источников показаны в Таблице 2. Совокупная точность работы системы по описанному алгоритму (процент правильно принятых решений) в процессе тестирования составляет 68,88%.

Источник	Число документов	Число вхождений неоднозначных терминов	Число правильных решений	Процент правильных решений
Известия	44	2525	1818	72.00
Ведомости	62	2697	1980	73.41
Независимая газета	42	2776	1846	66.50
Комсомольская правда	49	2240	1412	63.04
Всего	197	10238	7052	68.88

Таблица 2. Точность разрешения лексической многозначности по источникам публикаций

Заключение

В статье описана процедура разрешения лексической многозначности на основе Общественно-политического тезауруса. Такая постановка задачи относится к классу задач разрешения лексической многозначности в процессе сопоставления текстов с тезаурусами и онтологиями, разработанными в рамках конкретных предметных областей.

Полученный результат в целом соответствует результатам, показанным на конференции по разрешению многозначности SENSEVAL-3. Однако, мы полагаем, что для рассматриваемой постановки задачи за счет исключения из рассмотрения значительного количества частотных слов с большим количеством значений могут быть достигнуты более высокие результаты по точности выбора правильного значения. Поэтому мы рассматриваем проведенный эксперимент как начальный этап в серии разработки оптимального алгоритма по разрешению многозначности на основе Общественно-политического тезауруса.

Благодарности

Авторы благодарят Селиванову Т.М., Сидорова А.В., Чуйко Д.С., Штернова С.В. за вклад в проведение эксперимента.

Список литературы

1. Кобрицов Б.П. Методы снятия семантической многозначности // Научно-техническая информация, сер.2, N 2, 2004.
2. Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С. Нариньяни – М.: Наука, 2002. Т.2. С.338346.
3. Шелов С.Д. Термин. Терминологичность. Терминологические определения // СПб., 2003. – 280 с.

Agirre E., Rigau G. A Proposal for Word Sense Disambiguation using Conceptual Distance // Proceedings of the First International Conference on Recent Advances in NLP. Tzigov Chark, Bulgaria, September 1995.

4. Ide N. Veronis J. Word Sense Disambiguation: The State of the Art // Computational Linguistics, 1992. V.24. N1. P.1-40.

Jiang J. Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy // COLING 1997.

5. Kilgarriff A., Rosenzweig J. Framework and Results for English SENSEVAL // Computers and the Humanities, 2000. V34. P.15–48. (http://www.lexmasterclass.com/people/Publications/2000-KilgRosenzweig-Senseval1frame.pdf)
6. Magnini B., Cavaglia G. Integrating Subject Field Codes into WordNet // Proceedings of the Second International Conference on Language Resources and Evaluation LREC 2000, Athens, Greece, 2002.
7. Mihalcea R., Chklovsky T., Kilgarriff *A*. Framework and results for English SENSEVAL // SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, July 2004, Barcelona, Spain. 2004. P.25–28.

10. Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Five papers on WordNet // CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990.
11. Patwardhan, S., Banerjee, S., Pedersen, T. Using Measures of Semantic Relatedeness for Word Sense Disambiguation // CICLING 2003.
12. Resnik P. Using information content to evaluate semantic similarity // IJCAI 1995.
13. Snyder B., Palmer M. The English all-words task // Proceedings of SENSEVAL-3. Third International workshop on the Evaluation of Systems for the Semantic Analysis of Texts. 2004. P.41-43
14. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network // Proceedings of the International Conference on Information & Knowledge Management (CIKM), 2, 1993. P. 67-74.
15. Voorhees E. Using WordNet to disambiguate word senses for text retrieval // Proceedings of SIGIR-1993. 1993. P.171-180.
16. Vossen Piek, Rigau G., Alegria I., Agirre E., Farwell D., Fuentes M. Meaningful results for Information Retrieval in the MEANING project // Proceedings of Third International WordNet Conference, 2006.