Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Выравнивание синсетов тезаурусов YARN и PWN

Агапов Г.Д.
Научный руководитель: Фильченков А. А.
Консультант: Браславский П.И.

Санкт-Петербург
2016

ОГЛАВЛЕНИЕ

Стр.

ВВЕДЕНИЕ

В последние годы такая область знаний, как обработка естественного языка (Natural Language Processing или NLP) находит все больше и больше применений в различных сферах человеческой деятельности. К этой области в частности относятся задачи информационного поиска, извлечения информации, распознавания и синтеза речи, построения систем машинного перевода, вопросно-ответных систем, а также множество других задач, приложений.

Построения лингвистических систем “с нуля” — процесс крайне трудозатратный, требующий как правило усилий значительного числа специалистов достаточно широкого профиля. Основная сложность заключается в разработке и наполнении моделей, описывающих язык, позволяющих с ним работать. Именно поэтому при разработке лингвистическим систем обыкновенно прибегают к использованию готовых инструментов и ресурсов общего назначения, многие из которых есть в открытом доступе. Примерами таких инструментов и ресурсов могут служить: инструмент для извлечения морфологической информации MyStem [1], инструмент анализа лексики на основе моделей дистрибутивной семантики Word2Vec [2], открытый корпус текстов на русском языке OpenCorpora [3] и другие.

Наконец, широко распространенным видом ресурсов для работы с лексикой языка являются тезаурусы. С тех пор, как в 1990 году вышла первая версия тезауруса для английского языка Princeton WordNet (PWN), электронные тезаурусы нашли применение во многих приложениях NLP. Были разработаны аналогичные ресурсы для целого ряда языков, в том числе был предпринят ряд попыток построения тезауруса русского языка (РуТез, RussNet), последней из которых является YARN (Yet Another RussNet) — тезаурус современного русского языка, разработка которого началась в 2013 году.

Настоящая работа посвящана одной из задач, возникшей при построении тезауруса YARN, задачи выравнивания — сопоставления его понятий понятиям тезауруса PWN, т.е. нахождения для понятий лексики русского языка (хранящихся в YARN) соответствий в лексике английского языка. Подобная задача впервые решалась исследователями, работавшими над проектом EuroWordNet, и с тех пор является классической при построении тезаурусов

для новых языков. Ценность её решения заключается в возможности объединения тезаурусов различных языков в единую сеть, в которой, имея понятие одного языка можно будет легко получить доступ к соответствующим понятиям других языков. Как следствие, информацию о таких связях можно использовать в построении систем машинного перевода (для увеличения точности перевода), систем извлечения смысловой информации из текста и других.

В первой главе даются основные понятия, используемые в работе, проводится краткий обзор предметной области. Во второй главе формулируется задача выравнивания, рассматриваются предшествующие попытки её решения, ставятся цели, задачи, преследуемые авторами настоящей работы.

Во второй и третьей главах предлагается метод решения задачи выравнивания, условно разделенный на два этапа: автоматическое выравнивание (т.е. алгоритмический предпроцессинг) и выравнивание с применением техник краудсорсинга. В этих главах также приводятся результаты тестирования полученного метода.

ГЛАВА 1. **ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ**

В настоящей главе проводится краткий обзор предметной области. Даются определения основных понятий, связанных с тезаурусами. Рассматриваются ресурсы YARN и PWN, работе с которыми посвящена настоящая работа.

1.1. Электронные тезаурусы

Электронный тезаурус — лингвистический ресурс, представляющий собой семантическую сеть, описывающую лексику языка. В данной главе мы введем основные определения, связанные с тезаурусами и которые понадобятся нам в дальнейшем изложении. Подробнее о построении и применении тезаурусов можно прочесть в изданиях [4], [5].

1.1.1. Синсет

Основная единица, из которой состоят электронные тезаурусы — синсет (synset) от английских ***syn****onym* и ***set***, что можно перевести как набор синонимов. Синсет — объект, соответствующий понятию естественного языка, для которого составлен тезаурус. Синсет, как правило, содержит в себе следующую информацию:

- набор синонимов — слов, выражающих в лексическом наборе языка данное понятие
- определение — краткое описание сути понятия, соответствующего синсету
- примеры использования — предложения (фразы) языка, как правило демонстрирующие использование слов синсета в характеризующих синсет контекстах

В качестве примера рассмотрим синсет тезауруса Princeton WordNet (здесь и далее — PWN), описывающего английское понятие *[dog, domestic dog, Canis familiaris]* (соответствующее русскому понятию *[собака, пёс]* как животного):

Идентификатор, слова синсета (синонимический ряд):

SID-02086723-N [dog, domestic dog, Canis familiaris]

Определение:

> *a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds*

Пример использования:

– “the dog barked all night”

1.1.2. Связи в тезаурусах

Помимо синсетов тезаурус содержит информацию о связях между ними. Как между синсетами, так и между конкретными словами выделяются различные виды связей. Причем для слов разных частей речи (а равно и синсетов, которые из них состоят) характерны различные виды связей. Типы связей, характерные для различных частей речи подробно рассмотрены (на примере английского языка) создателями PWN в работах [6], [7], [8]. Мы лишь приведем в пример некоторые из них (на которые будем ссылаться в дальнейшем изложении материала).

Гиперонимия — отношение между понятиями, характеризуемое предикатом “A —- частный случай (частная форма, проявление) B”. При этом A называется гиперонимом, B — гипонимом. Например *дерево – дуб*, *дерево* — гипероним, *дуб* — гипоним.

Меронимия (холонимия) — отношение между понятиями, характеризуемое предикатом “A — (составляющая) часть B”. При этом A называется меронимом, B — холонимом. Например *дерево – ствол*, *дерево* — холоним, *ствол* — мероним.

Определения отношений синонимиии, антонимии дать несколько сложнее (существует несколько определений, опирающихся на различные соображения), мы будем полагаться на интуитивное понимание их читателем. Отметим лишь, что в отличие от изложенных выше, синонимия и антонимия — отношения между словами языка, а не понятиями.

1.1.3. Полисемия

Важно также дать определение понятию полисемии, которое не раз встретится в ходе последующего изложения. Полисемией (многозначностью) на-

зывают свойство языка, заключающееся в наличии у слова более чем одного значения. Например слово *лук* русского языка обозначает одновременно и приспособление для стрельбы, и овощ. Это два принципиально различных понятия. Слова, относящиеся к более, чем одному понятию, мы будем называть полисемичными.

С точки зрения тезаурусов, полисемичность языка означает, что одно и тоже слово может встречаться в более, чем одном синсете. И более того, могут встречаться синсеты с одними и теми же наборами синонимов, в связи с чем введем также понятие полисемии на уровне синонимических рядов: будем называть синонимический ряд (набор слов-синонимов) полисемичным, если существует более одного понятия в языке, характеризуемых этим рядом.

В качестве примера полисемичного набора синонимов рассмотрим английский *[force]*. В PWN существует целых три синсета, полностью ему соответствующих:

- SID-05201846-N [force] (a powerful effect or influence)
- SID-11479041-N [force] ((physics) the influence that produces a change in a physical quantity)
- SID-08224784-N [force] (a group of people having the power of effective action)

Различаются синсеты с совпадающими наборами синонимов как правило по определениям, примерам использованим, а также связям с другими синсетами тезауруса.

1.2. Princeton WordNet

Princeton WordNet (PWN) [5] — первый проект построения электронного тезауруса. Разработка тезауруса ведется с 1985 года группой исследователей из Принстонского университета. Авторами была поставлена цель построить лексический ресурс, тезаурус, со структурой, аналогичной структуре, в которую организуются знания о языке в человеческом сознании (в соответствии с актуальными на тот момент психолингвистическими теориями).

Исследователями была проделана работа по систематизации знаний об устройстве лексики английского языка, отношениях между понятиями и от-

дельными словами. По результатам этой работы были разработаны форматы для нового тезауруса, началось его заполнение.

На данный момент актуальной версией PWN является версия 3.1. Тезаурус содержит более 100 тысяч английских понятий, работа над его пополнением (и уточнением) продолжается и сейчас.

Проект PWN имеет открытую лицензию и доступен для использования как в исследовательских, так и в коммерческих целях.

1.3. Тезаурусы русского языка

На момент написание настоящей работы, известно как минимум три проекта создания электронного тезауруса для русского языка:

– RussNet
– РуТез
– Yet Another RussNet (YARN)

Тезаурус RussNet [9] разрабатывается с 1999 года силами исследователей из Санкт-Петербургского государственного университета.

Проект РуТез [4] представляет из себя вообще говоря не тезаурус типа WordNet, но лингвистическую онтологию. Обсуждение понятия лингвистической онтологии выходит за рамки настоящей работы, ознакомиться с ним можно, например, на странице описания проекта РуТез [10]. Проект РуТез имеет представленние в формате тезауруса, RuWordNet [11]. Разрабатывается в Московском государственном университете с 2002 года.

Результаты проекта RussNet недоступны для использования, в настоящий момент ведется активная работа по систематизации, форматированию полученных результатов для дальнейшей публикации.

Проект РуТез доступен только для некоммерческого использования и в урезанной версии (РуТез-lite). Актуальную версию можно получить, связавшись с автором.

1.4. Yet Another RussNet

Yet Another RussNet (YARN) [12] — проект создания нового открытого электронного тезауруса русского языка. Разрабатывается с 2013 года представителями нескольких российских университетов.

В настоящий момент тезаурус находится в стадии активной разработки. Характерной особенностью тезауруса является активное применение в его построении техник краудсорсинга. В частности разработан интерфейс, посредством которого любой желающий может поучаствовать в его наполнении и редактировании.

К сожалению, тезаурус в текущем состоянии обладает рядом существенных недостатков (работа над устранением которых активно ведется):

- неполнота покрытия — многие, в том числе достаточно часто используемые понятия русского языка в тезаурусе отутствуют
- наличие определений у менее, чем 5% синсетов
- отсутствие связей между синсетами
- наличие синсетов, в которых использованы слова различных понятий (вследствие, как правило, включения слов понятий, связанных с данным отношениями гиперонимии/меронимии)
- наличие дубликатов — групп из более, чем одного синсета, относящихся к одному и тому же понятию русского языка

Все эти недостатки будут нами учтены в построении метода выравнивания (подробнее каждая из этих проблем будет рассмотрена в главе 3).

Резюме

В изложенной главе были даны определения основных понятий, используемых в настоящей работе, в частности понятий тезауруса, синсета, гиперонимии, меронимии, полисемии.

Дан короткий обзор проекта Princeton WordNet, существующих тезаурусов русского языка, а также интересующего нас главным образом тезауруса YARN.

ГЛАВА 2. **ПОСТАНОВКА ЗАДАЧИ**

В данной главе формулируется задача выравнивания тезаурусов, рассматриваются предшествовавшие попытки её решения в проектах EuroWordNet, BalkaNet. Обсуждается задача выравнивание синсетов тезаурусов YARN и PWN, определяются цели настоящей работы.

2.1. Задача выравнивания

В главе 1 было дано определение синсета, как основной единицы тезауруса, объекта, описывающего понятие естественного языка, для которого построен тезаурус. Очевидным наблюдением, касающимся понятий разных языков, является факт нахождения в них слов, понятий, описывающих одни и те же объекты, явления, свойства реального мира. Например, для человека, сколько-нибудь владеющего русским и английским языками, будет очевидно, что английское понятие *[dog, domestic dog]* описывает приблизительно тот же объект, что и русское *[собака, пёс]*.

Как следствие, можно сформулировать гипотезу о том, что можно выделить некоторое множество достаточно распространенных, присутствующие в обоих языках понятий, и, взяв тезаурусы, построенные для этих языков, объединить такие понятия соответствующим типом связей (между тезаурусами). Множество таких общих понятий принято обозначать как ILI (Interlingual index, межязыковой индекс). Под задачей *выравнивания* тезаурусов A, B понимают объединение задач выделения ILI — множества понятий, присутствующего в каждом из двух тезаурусов, сопоставления понятиям A из ILI соответствующих им понятий B.

2.2. Предшествующие работы

Впервые задача выравнивания была сформулирована в рамках работы над проектом EuroWordNet [13] — попыткой создания тезаурусов для группы европейских языков (датского, итальянского, испанского, немецкого, французского, чешского и эстонского). Авторами было введено понятие ILI как неструктурированного набора понятий, собранного с единственной целью

предоставления возможности эффективного сопоставления синсетов различных языков. ILI представлял из себя список из 1024 идентификаторов PWN 1.5, так называемых основных понятий (Base concepts или BCs).

Для каждого из языков тезаурус создавался независимо. Общие понятия выделелись на основе специально созданной онтологии (Top ontology), состоявшей из 63 семантических критериев, и кроме того, полученные понятия проверялись на наличие в тезаурусах всех рассматриваемых языков. Подробнее с проведением выравнивания в проекте EuroWordNet можно ознакомиться в работе [14].

Другим масштабным проектом создания группы тезаурусов, ставившим своей целью проведение процедуры выравнивания, является проект BalkaNet [15] — проект создания тезаурусов для языков балканского региона: болгарского, чешского, греческого, румынского, турецкого, сербского. В своей работе авторы использовали опыт проекта EuroWordNet: также независимо строились тезаурусы, а уже впоследствии проводилось их выравнивание с PWN. Для построения ILI авторами были взяты 1024 BCs проекта EuroWordNet, которые были впоследствии дополнены новыми понятиями, в результате чего получился набор из 4680 BCs, по которым проводилось выравнивание полученных тезаурусов.

Заметим, что в обоих проектах и построение тезаурусов, и проведение выравнивания проводилось вручную (с использованием некоторых вспомогательных программных средств) силами работающих параллельно команд лингвистов.

2.3. Выравнивание синсетов YARN и PWN

Задача настоящей работы заключается в выравнивании синсетов тезаурусов YARN и PWN.

Цель настоящей работы — построить метод проведения выравнивания большинства синсетов из BCs без привлечения экспертов, с помощью автоматического предпроцессинга и применения техник краудсорсинга.

Данный подход представляется авторам более перспективным, так как при условии достижения достаточного высокой надежности получаемых связей позволит выполнять выравнивание не только для заранее определенного

(фиксированного) множества основных понятий, но и для произвольных понятий при условии наличия таковых в обоих языках.

В качестве множества синсетов для тестирования метода принято подмножество основных понятий-существительных (3143 синсета), выделенных в рамках проекта BalkaNet.

Резюме

В данной главе была сформулирована задача выравнивания синсетов двух тезаурусов, рассмотрены некоторые из предшествующих попыток решения задач данного типа. Сформулированны задача, цели настоящей работы.

ГЛАВА 3. **АВТОМАТИЧЕСКИЙ ПРЕДПРОЦЕССИНГ**

В данной главе рассматривается применение автоматического подхода для разрешения задачи выравнивания.

Под автоматическим подходом здесь и далее понимается попытка алгоритмическим путём выделить для данного синсета из PWN как можно меньшее множество синсетов из YARN, содержащее как элементы синсеты, наилучшим образом описывающее исходное понятие из PWN (далее — множество кандидатов).

3.1. Граф связности

В данном разделе мы введем понятие графа связности — структуры, используемой для нахождения множества кандидатов данного синсета.

3.1.1. Наивный подход

Наивный подход к решению задачи выравнивания тезаурусов для языков *A*, *B* заключается в следующем:

- Рассмотрим все синсеты S_A тезауруса для языка A
- Для каждого синсета S_A рассмотрим множество слов w, содержащихся в нём
- Для каждого слова w с помощью переводного словаря найдём множество его переводов $T_{AB}(w)$ — соответвующих ему слов языка B
- Для каждого слова $v \in T_{AB}(w)$ найдем множество синсетов $S_B(v)$ тезаурса для языка B, содержащих его в синонимическом ряду
- Объединением $C(S_A) = \bigcup_{\forall w \in S_A} \bigcup_{\forall v \in T_{AB}(w)} S_B(v)$ всех таких множеств получаем искомое множество кандидатов

Однако применение такого подхода непосредственно приводит к непозволительно большому размеру полученного множества $C(S_A)$, элементы которого зачастую лишь отдаленно связаны с исходным понятием S_A. Например, для синсета из PWN *SID-02116752-N {wolf}* с определением “any of various predatory carnivorous canine mammals of North America and Eurasia that usually hunt in packs” было найдено 47 синсетов из YARN.

Произошло это вследствие многозначности слова “wolf” английского языка; существуют как минимум три понятия, содержащие это слово:

– волк (хищное животное) — *SID-02116752-N [wolf]*
– жестокий человек, чудовище — *SID-09864997-N [beast, wolf, savage, brute, wildcat]*
– соблазнитель, ловелас – *SID-10806873-N [wolf, woman chaser, skirt chaser, masher]*

Соответственно, в переводном словаре для каждого из этих трех смыслов будут соответствующие вхождения слов русского языка. Кроме того, полисемично и русское слово “волк”, которое относится как ко всем трем указанным выше понятиям, так и к еще одному:

– нелюдим, волк, бирюк

Соответственно в сформированном множестве кандидатов будут синсеты, относящиеся как ко всем четырём понятиям, содержащим слово волк, так и не имеющие к волкам практически никакого отношения как, например, *s34913 {урод, чудовище}*.

3.1.2. Граф связности

Введем понятия графа связности.

Граф связности двух синсетов A и B — двудольный ориентированный взвешенный граф $\langle A, B, W \rangle$:

– Вершины - синсеты тезаурусов
– Одна доля содержит все синсеты тезаурса A, другая - все синсеты тезауруса B
– Вес ребра - вещественное число $w \in [0, 1]$, имеющее смысл значительности связи между двумя синсетами
  – 0 — синсеты описывают никак не связанные понятия
  – 1 — синсеты описывают одно и то же понятие

Будем считать граф связности полным (однако лишь небольшое количество весов отличны от нуля). Ориентированность введена из соображений удобства: строить граф мы будем преимущественно посредством перевода данных синсета одного тезауруса в язык другого, ребро $a \to b$ ($a \in A, b \in B$) будет соответствовать результату перевода синсета a в тезаурус языка B.

В дальнейшем под ребром, весом в зависимости от контекста мы будем понимать как ориентированные ребра (веса), так и неориентированные (полученные усреднением весов двух взаимнообратных ориентированных).

Вернемся теперь к решаемой нами задаче. Предположим, мы каким-то образом умеем строить граф связности, веса которого достаточно хорошо показывают связь между синсетами двух языков. Тогда задачу выравнивания решить сравнительно просто:

– Преобразуем граф связей в неориентированный (например, взяв среднее от весов обоих ребер между двумя синсетами)
– Для данного синсета $a \in A$ сортируем связанные с ним синсеты $b \in B$ по убыванию весов
– Из отсортированного списка синсетов выберем первые $\leqslant K$ элементов — множество кандидатов

Собственно, проблема выравнивания теперь может быть переформулирована как проблема построения графа с весами, релевантными связям между понятиями языков данных нам тезаурусов.

3.2. Применение меры Жаккара

Итак, мы хотим построить метрику, достаточно хорошо показывающую связи между понятиями, имеющими место в натуральном языке. За основу для такой метрики кажется естественным взять т.н. меру Жаккара.

3.2.1. Определение

Мера Жаккара (Jaccard index) для двух произвольных множеств A, B определяется как отношение мощности пересечения к мощности объединения:

$$JI(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

При решении задачи мы использовали модификацию меры Жаккара, отличающуюся несколько большими значениями весов:

$$JI'(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

3.2.2. Применение

Применить меру Жаккара для связывания синсетов из разных языков “напрямую” очевидно невозможно, т.к. слова разных языков гарантированно будут различны (и мы получим $JI'(a,b) \equiv 0$). Т.е. нам нужно как минимум решить задачу сопоставления синсету $a \in A$ из тезауруса языка A множества $\tau(a)$ наборов слов языка B, и уже эти наборы можно будет сравнить по мере Жаккара с неким синсетом $b \in B$.

Задать $\tau(a)$ можно множеством способов. Например, следующим образом:

– переводить каждое слово $\alpha \in a$ синсета $a \in A$
– $\tau(a)$ определить как множество, содержащее единственным своим элементом объединение всех переводов всех α

Такой подход, однако, на практике оказывается неприменим, т.к. в силу полисемии, синонимии каждому слову как правило соответствует сразу несколько переводов, $x \in \tau(a)$ окажется слишком раздутым, что приведет к низким и малорелевантным значениям JI'.

3.2.3. Структура словарей

Рассмотрим структуру классического двуязычного (переводного) словаря. Словарь разбит на статьи, каждая соответствует одному слову языка A.

Существенной особенностью большинства словарей (замеченной в частности и авторами работы [16]) является то, что в пределах статьи слова даны не просто в виде списка, но в сгруппированном по значениям виде.

Например, так выглядит статья для слова *окружение* в словаре LingvoUniversal (версия 2003 года):

окружение

а) environment; surroundings; milieu
б) environs, nighbo(u)rhood; entourage
в) (воен.) encirclement

Заметим, что каждая такая группа в словарной статье предствляет собой набор слов из языка B (как правило синонимов), потому:

– $\tau(a)$ можно определить как объединение всех таких групп для всех слов исходного синсета $a \in A$
– такое построение представляется достаточно естественным, т.к. мы будем сравнивать синсеты как наборы синонимов с переводами — другими наборами синонимов

3.2.4. Метрика

Построим граф связности, используя в качестве метрики меру Жаккара, построенную поверх определенного описанным выше способом $\tau(a)$. Вес ребра между синсетами $a \in A$, $b \in B$ определим как:

$$w(a \in A, b \in B) = max\{JC'(x, b) \mid x \in \tau(a)\}$$

Полученный таким способом граф, однако, обладает рядом недостатков:
– по-прежнему непозволительно большое число связей со схожими (достаточно высокими) весами
  – для синсета *SID-13957629-N [environment] (среда, контекст, окружение)* — 22 связи
  – для синсета *SID-08637195-N [public square, square] ((городская) площадь)* — 35 связей
  – для синсета *SID-08016141-N [set] (математическое множество)* — более 50 связей
– далеко не всегда связи устроены таким образом, что в первых 5-10 синсетах лежит искомый синсет
  – из-за проблемы дубликатов (подробнее — см. следующий раздел) часто 7-10 синсетов с максимальным весом соответствуют одному нерелевантному понятию русского языка, а то понятие, которое в действительности нас интересует находится на 15-17 месте
– для заданного синсета из PWN веса синсетов из YARN чаще всего группируются по значениям, как 1, ± 0.7, ± 0.35 и между собой малоразличимы

Кроме того было замечено, что для различных синсетов из PWN значения максимального веса смежных ребер значительно разнятся, от 0.3 до 1, из-за чего первоначально выдвинутая нами гипотеза отсекать по порогу (threshold) оказалась неприменимой. Однако если рассматривать конкретно взятый син-

сет из PWN, найденные синсеты из YARN вообще говоря неплохо различаются между собой, потому дальнейшее использование меры Жаккара показалось нам целесообразным (после применения к ней ряда улучшений).

3.3. Основные проблемы, их решение

В данном разделе будут рассмотрены основные проблемы, с которыми мы столкнулись, применив описанную в предыдущем разделе метрику, решения (улучшения), применённые к метрике, позволившие уменьшить эффект, производимый некоторыми из этих проблем.

Будут рассмотрены некоторые проблемы, уже упомянутые в разделе 1.4 и специфичные для YARN:

– синсеты-дубликаты
– многопонятийные/смешанные синсеты
– неполнота YARN

Также и некоторые, характерные для задачи в целом:

– длинные синсеты
– полисемия

3.3.1. Синсеты-дубликаты

Как было упомянуто в предыдущем разделе, после применения метрики на основе меры Жаккара, количество полученных связей оказалось непозволительно большим. Основным фактором, приведшим к такому результату, были именно синсеты-дубликаты. Напомним, дубликатами мы называем синсеты, описывающие одно и то же понятие естественного языка.

Чтобы продемонстрировать масштаб проблемы, рассмотрим множество синсетов, связанных с *SID-10204565-N [hope] (someone (or something) on which expectations are centered)* с отсечением по весу 0.2 (синсеты сгруппированы по понятиям):

– надежда как возможность в будущем
  – s1816 [возможность, допустимость, вероятность, случай, надежда]
  – s277 [перспектива, грядущее, будущее, будущность, надежда, шанс, ожидание]

    - s30159 [мечта, надежда, ожидание, чаяние]
    - s30630 [надежда, ожидание, упование]
    - s30631 [надежда, упование]
    - s30632 [надежда, упование, чаяние]
- ложная надежда, (само)обман
    - s22799 [воздушные замки, ложное представление, видимость, заблуждение, иллюзия, надежда, обман, химера]
- надежда как способ пережить проблему
    - s13028 [надежда, прибежище, утешение]
- надежда как опора
    - s30628 [надежда, надёжа, оплот]
    - s30629 [надежда, надёжа, опора]
    - s6550 [надежда, основание]

Всего 11 связей, что вообще приемлимо немного (10-15 синсетов можно просмотреть вручную). Но все-таки кажется разумным сперва “схлопнуть” синсеты, описывающие одно понятие в один синсет, а уже потом пытаться связать этот объединенный синсет с синсетом из PWN. Кроме того, для подавляющего числа других синсетов из BCs количество связей значительно превышает 11, и без отсечения дубликатов дальнейший процессинг попросту невозможен.

Заметим, что “схлапывание” синсетов — вообще говоря отдельная задача, непосредственно не связанная с задачей выравнивания. Как уже было отмечено, YARN находится в стадии активной разработки, и исследование подходов решения проблемы дубликатов — одно из основных направлений работы.

Поскольку редактирование существующих синсетов выходит за рамки решаемой задачи, было принято решение не объединять синсеты с одним смыслом, но группировать, воспользовавшись некоторой разумной эвристикой для классификации двух синсетов как описывающих одно понятие языка. В частности в работе [17] для нахождения потенциальных дубликатов предлагается рассмотреть все пары синсетов, пересекающиеся по двум и более словам. Авторы, ссылаясь на работу [18], утверждают, что большинство таких пар являются парами из двух синсетов-дубликатов.

Именно идея объединения синсетов, пересекающихся хотя бы по двум словам, была взята нами за основу эвристики. После нахождения для данно-

го синсета из PWN множества кандидатов из YARN, мы жадно объединяем кандидатов в смысловые группы — кластеры. Кроме, собственно, эвристики пересечения по двум, будем использовать следующие дополнительные соображения:

– перед сравнением разумно пропустить синсеты через стеммер, т.к. в YARN встречаются синсеты, в которых слова даны, например, в форме множественного числа
– если два одноэлементных синсета полностью совпадают, их тоже будем считать дубликатами.

Сформулируем алгоритм формирования кластеров (повторяем шаги а)–в), пока множество нерасмотренных синсетов не станет пустым):

а) берем любой еще не рассмотренный нами синсет
б) создаём из него одноэлементный кластер
в) последовательно перебираем множество нерассмотренных синсетов
    – добавляем в кластер те, которые пересекаются по обозначенному выше критерию с хотя бы одним элементома кластера (синсетом, добавленным на предыдущем шаге)
    – синсеты, добавляющиеся в кластер мы исключаем из дальнейшего рассмотрения

Для каждого класса мы определяем ровно один синсет-представитель, в дальнейшем будем рассматривать только связи между исходным синсетом из PWN и представителями кластеров. В качестве представителя выбирается синсет с наибольшим весом.

На рисунке 3.1 изображен подграф для *SID-10204565-N [hope]*, разбитый на кластеры. Синсеты были объединены в кластеры согласно разбиению по смыслам, использованном в списке выше за исключением синсета *s6550 [надежда, основание]*, который был выделен в отдельный одноэлементный кластер.

Разбиение на кластеры значительно улучшило свойства графа связей:

– существенно уменьшилось число связей
– для подавляющего числа синсетов из PWN наилучший кандидат на связывание стал находиться среди первых 15 синсетов

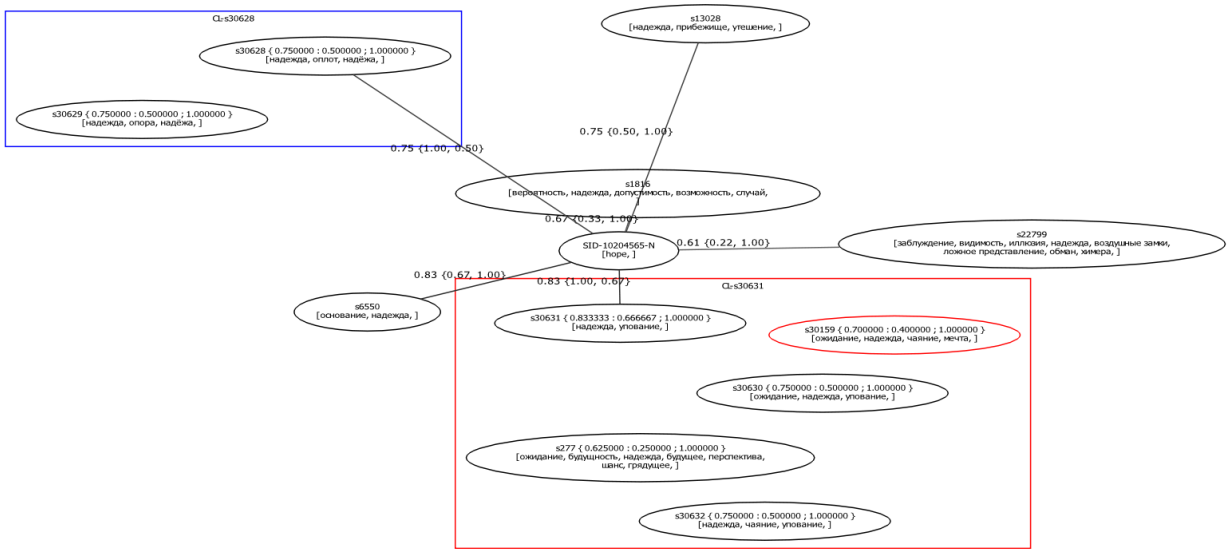

Рис. 3.1 — Кластеризованный граф связей для SID-10204565-N (hope)

3.3.2. Полисемия

При переводе мы рассматриваем все слова исходного синсета, и если какое-нибудь из них полисемично, то в графе каждый смысл этого слова потенциально индуцирует по ребру (а вероятно и не по одному). Понятно, что среди полученных связей будет достаточно избытычных, с синсетами из YARN, не относящимися к исходному синсету. В идеале нам хотелось бы опеределить метрику таким образом, чтобы избыточные связи обладали меньшим весом, чем связи с синсетами YARN, близкими по смыслу к исходному.

В общем случае разрешить полисемию можно только посредством глубокого анализа определений синсетов, тезаурусных связей и т.д. Провести такой анализ автоматически представляются достаточно трудной задачей. В случае с YARN в силу отсутствия тезаурусных связей, определений (для большинства синсетов) такой анализ и вовсе невозможен, потому проблему полисемии мы будем решать посредством применения краудсорсинга (о чем будет рассказано в главе 4).

В следующих подразделах мы рассмотрим несколько частных случаев проявления полисемии, для которых посредством изменения метрики получилось существенно улучшить её релевантность.

3.3.3. Длинные и смешанные синсеты

Длинным синсетом мы условно называем синсет, состоящий из более чем двух-трёх слов. Смешанным — синсет, включающий в себя более одного понятия естественного языка. Понятно, что в тезаурусе в идеале не должно быть смешанных синсетов, и каждому синсету должно соответствовать ровно одно понятие естественного языка.

Однако в действительности некоторые синсеты могут быть настолько близки по смыслу друг к другу, что даже эксперту-лингвисту сложно будет однозначно определить, к одному они относятся понятию или к разным (и, соответственно, если объединить, сложно определить, является ли синсет смешанным (определяющим два близких понятия)). Подробнее об этом и других затруднениях, возникающих при построении тезаурусов можно прочесть в [19].

Кроме того, YARN является ресурсом в стадии активной разработки, вследствие чего в нем значительно чаще встречаются синсеты, содержащие слова более одного понятия.

Смешанный синсет может и не являться длинным, однако для YARN это наблюдение зачастую выполняется. Мы рассмотрим случай, когда длинный синсет (тем более если он и смешанный) почти наверняка существенно ухудшит результат применения меры Жаккара, предложим способ нивелировать этот эффект.

При задании веса в формуле 3.2.4 мы использовали максимум, причем максимум брался по всем переводным группам всех слов исходного синсета языка A. Заметим, что если в синсете есть слово α, относящееся к более, чем одному понятию, то в случае, когда его перевод близок по Жаккару к синсету $b \in B$ (т.е. значение $w = max\{JC'(x, a) \mid x \in \tau(\{\alpha\})\}$ достаточно высоко), между a и b будет ребро веса $\leqslant w$, даже в том случае, когда α — единственное слово a, имеющее переводную связь с b. Для небольших синсетов это не представляет существенных затруднений, однако чем больше слов в синсете, тем вероятнее появление такого рода избыточных связей.

Из формулировки проблемы следует и идея её решения — накладывать штраф на связь, если менее μ% слов синсета участвуют в ней. В качестве функции штрафа удобно выбрать такую, которая при небольших отклонениях $h < \mu$ не вносила бы существенного вклада в веса, однако чтобы с умень-

шением h множитель штрафа уменьшался сильнее, в пределе ($h \to 0$) к 0. Из этих соображений функция штрафа была выбрана на основе функции нормального распределения:

$$m(h \mid \mu, \sigma) = \begin{cases} \frac{f(h)}{f(\mu)}, & h < \mu \\ 1, & \text{иначе} \end{cases}$$

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

В качестве параметров распределения были взяты следующие значения:

– матожидание: $\mu = 0.5$, т.е. штраф накладывается только в том случае, когда менее половины слов синсета a связаны с b
– стандартное отклонение: $\sigma = 0.2$

3.3.4. Полисемия на уровне синонимических рядов

Применительно к задаче связывания полисемия на уровне рядов порождает ситуацию, когда нам не всегда достаточно анализа синсетов как множеств синонимов и их переводов, но желательно также учитывать определения, указанные в синсетах.

В разделе 1.1.3 в качестве примера нами рассматривались синсеты PWN, состоящие из единственного слова *force*:

– SID-05201846-N [force] (a powerful effect or influence)
– SID-11479041-N [force] ((physics) the influence that produces a change in a physical quantity)
– SID-08224784-N [force] (a group of people having the power of effective action)

Аналогично, в YARN существует 7 синсетов с наборами слов *[сила]*, каждый из них оснащен своим определением. Строя подграф для, например, *SID-11479041-N* мы получим кластер из 7 этих синсетов, связанный с *SID-11479041-N* ребром веса 1, причем представитель будет выбран произвольным образом (тогда как разумнее было бы выбрать именно тот, который обладает значением силы как термина из физики).

Однако учет определений имеет даже больший потенциал применений. Предположим, наборы синонимов в переводе у нас тоже будут оснащены краткими определениями, которые будут его отличать от других наборов

в переводе данного слова. Например для переводов слова *force* (данные из Wiktionary):

а) [сила] (physical quantity that denotes ability to accelerate a body)

б) [сила, мощь, дурь] (anything that is able to make a big change in person or thing)

в) [власть] (law: legal validity)

г) [сила, насилие] (law: unlawful violence or lawful compulsion)

Сравнив определение из первого перевода с определением синсета из PWN, мы поймем, что из четырёх предложенных переводов нам интересен именно первый.

Сравнивать определения синсетов двух связываемых тезаурусов можно было бы, например, применяя машинный перевод к одному из определений, сравнивая затем получившиеся предложения как последовательности слов. Однако мы посчитали, что реализовывать это нецелесообразно, т.к. на практике определениями снабжено менее 5% синсетов YARN.

Покрытие определениями переводов в используемом нами словаре Wiktionary значительно шире. Кроме того, определения даны на том же языке, что и определения исходного (переводимого) синсета, что позволяет сравнивать их как последовательности слов и без применения машинного перевода.

Сравнивать определения как последовательности слов можно различными способами. Однако заметим:

– перед нами не стоит задача точного сравнения определений, нам нужно только различать одни переводы от других
– определения в Wiktionary как правило короче определений из PWN, содержат несколько ключевых слов, которые как правило отсутствуют в определениях других переводов

В силу соображений, данных выше, мы нашли целесообразным использовать следующий простой подход:

– преобразуем определения перевода t, исходного синсета a в множества слов $ws(t)$, $ws(a)$
  – нормализуем формы слов с помощью стеммера
– за метрику похожести определений возьмём $s(t, a) = \frac{ws(t) \cap ws(a)}{ws(t)}$

Функция веса ребра между синсетами a, b тогда примет следующий вид:

$$w(a \in A, b \in B) = max\{p \cdot JC'(t,b) + (1-p) \cdot s(t,b) \mid t \in \tau(a)\}$$

Где p - некий фиксированный коэфициент. Нами был использован $p = 0.7$.

3.4. Дополнительные улучшения

В данном разделе мы рассмотрим ряд улучшений, не рассмотренных в предыдущем разделе, но ощутимо улучшивших характеристики графа связей.

3.4.1. Используемые отсечения

Первоначально разработанную метрику предполагась рассмтривать, как непосредственный способ оценки близости синсета из PWN и YARN. Предполагалось, что значение, близкое к единице будет указывать на близость синсетов, значения ниже $T = 0.3$ — отсутствие значительной связи между синсетами. И как следствие, планировалось получить множество кандидатов удалением всех ребер, меньших T.

Однако, впоследствии было замечено, что для некоторых синсетов вес наиближайшего по метрике кандидата не привышает 0.35, причем среди кандидатов есть релевантные. Причем, что также хорошо наблюдалось, между собой кандидаты на связывание достаточно хорошо отличались по весам, что позволяло выбрать из них первые 10-15, будучи уверенными, что кандидаты с меньшим весом наверняка не представляют для нас интереса. Кроме того, поскольку мы строили граф связей как ориентированный граф, имело смысл рассматривать не только средний (неориентированный) вес, но и веса обоих полурёбер.

Именно поэтому было принято отсекать лишние связи по комбинированному критерию: выбирать первые $N_{max} = 15$ кандидатов, вес полуребер которых больше $T' = 0.2$, средний вес — больше $T = 0.2$.

3.4.2. Использование словарей

Когда работа над процедурой связывания только началась, мы использовали только один (оффлайновый) словарь — Lingvo Universal (версия ок. 2003 года), доступный в формате Stardict [20]. Одним из первых наблюдений для нас был явный недостаток покрытия словаря: некоторые слова, не столь редкие в употреблении либо в нем, либо в статьях перевода отстутствовали, тем самым затрудняя нахождение всех релевантных кандидатов.

Следующим шагом мы попробовали использовать онлайн сервис, Яндекс.Словари. В этом сервисе также используются словари Lingvo (только уже последней версии). Данный сервис обладал ощутимо большим покрытием, однако имел два существенных недостатка: время, требуемое на исполнение HTTP запроса и суточное ограничение на число запросов.

Процедура выравнивания устроена таким образом, что для того, чтобы установить связь между двумя синсетами $a \in A$ и $b \in B$, нам нужно сперва перевести все слова a в язык тезауруса B, а затем для всех найденных синсетов из B произвести обратный перевод в язык A. Допустим, a состоит из трех слов и для a нашлись переводы в 5 синсетов B, а каждый из этих пяти также состоит из трёх слов. Итого получаем $3+5\cdot3 = 18$ запросов к онлайн-сервису. На практике из-за многозначности многих слов, количество запросов для одного синсета может достигать и значительно больших значений (более ста - не редкость), вследствие чего лимит сервиса (10000 запросов) исчерпывается за первые 100-200 синсетов PWN.

Именно поэтому было решено использовать:

– кэширование запросов к онлайн-сервису
– обращаться к онлайн сервису только в случае, если перевод не найден при помощи оффлайнового словаря

Далее нами было опробовано использование результатов парсинга ресурса Wiktionary [21] в качестве дополнительного словаря. Ресурс оказался чрезвычайно ценным, обладающим значительной полнотой покрытия, точностью, и кроме того для большого числа синсетов в Wiktionary обнаружились описания на языке исходного синсета, использование которых помогло нам значительно улучшить меру (см. раздел 3.3.4).

Чтобы максимально полно и эффективно использовать данные всех этих словарей, для каждого направления перевода (англо-русский и

русско-английский) они были объединены в две конфигурации: полную и ступенчатую.

Полная конфигурация объединяет выдачи от всех словарей, ступенчатая же состоит из нескольких ступеней, словари следующей используются только если слово не было найдено в выдачах словарей предыдущей ступени.

Разделение на две конфигурации было в введено со следующей целью: во время тестирования мы столкнулись с проблемой *синсетов-сирот,* т.е. синсетов, для которых не было найдено ни одного кандидата из YARN. Причем таких синсетов обнаруживалось достаточно много, в то время как для многих переводы в словарях находились. В процессе разбирательства мы обнаружили, что для многих из этих синсетов по переводу находились синсеты из YARN, но не все, и именно те, которые могли образовать связи по обратным переводам, не находились. Причем, что мы также заметили, нередко использование полной конфигурации решало проблему (однако как было отмечено, излишне загружать онлайн-сервис запросами нам бы тоже не хотелось).

Решение было найдено следующее: поиск кандидатов проводился в три этапа, с каждым ослаблялись критерии отсечения, на втором и третьем этапах использовалась полная конфигурация. Применение такого улучшения позволило найти кандидатов (притом как правило релевантные синсеты) для дополнительных 8% тестовой выборки.

3.5. Тестирование

Предложенный метод построения графа связей был протестирован на множестве из 100 синсетов из PWN. Были сформированы множества кандидатов. Для 97 синсетов нашелся хотя бы один кандидат.

Для каждого синсета было сформированно задание из первых 15 кандидатов. Эксперту-лингвисту предлагалось из этих пятнадцати выбрать несколько (от 1 до 4) кандидатов, наиболее подходящих исходному синсету. Кроме того, некоторые из отмеченных синсетов можно было пометить, как относящиеся к понятию русского языка, полностью идентичного английскому понятию (см. скриншот 3.2). Данные, полученные от эксперта были аккуратным образом обработаны, ответы для некоторых синсетов были подкорректированы

(во взаимодействии с экспертом), в результате чего была получена эталонная разметка для тестируемого множества из 100 синсетов.

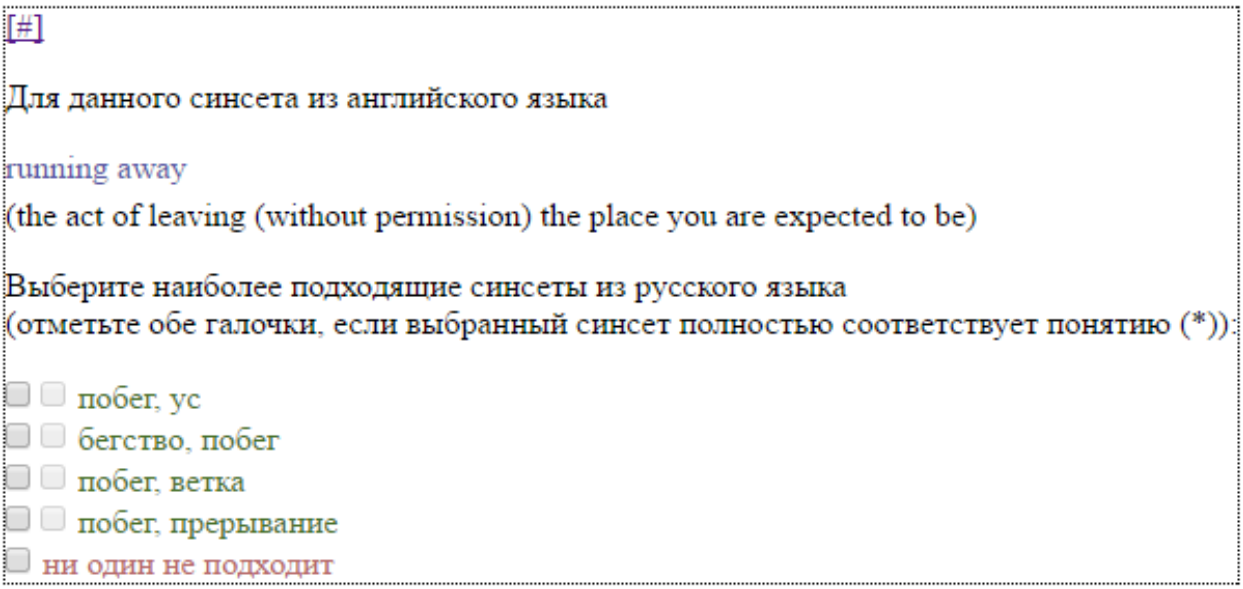

Рис. 3.2 — Интерфейс разметки эксперта

В таблице 3.1 представлен анализ результатов разметки. В первой колонке — порядковый номер кандидата, k, во второй — для скольки синсетов первый кандидат, отмеченный экспертом — k-й в списке кандидатов. Третья колонка — для скольки синсетов первый кандидат, отмеченный экспертом и как близкий, и как полностью соответствующий — k-й в списке кандидатов.

Таблица 3.1 — Анализ результатов разметки.

Кандидат	Кол-во (все)	Кол-во (полное соответствие)
Всего	86	73
$k = 1$	65	56
$2 \leqslant k \leqslant 5$	14	13
$6 \leqslant k \leqslant 10$	5	4
$11 \leqslant k \leqslant 15$	2	0

Получается 11 синсетов, для которых в первых пятнадцати кандидатах не нашлось соответствия. Эти 11 синсетов были просмотрены вручную. Для двух синсетов в действительности существовали кандидаты на привязку, но

были помещены в кластер с нерелевантным представителем (эвристика сработала слишком агрессивно).

Для остальных 9 синсетов действительно не было найдено соответствия (что означает либо недостаточно полный перевод, либо отсутствие подходящего синсета в YARN вообще). С полученным распределением можно ознакомиться в таблице 3.2.

Таблица 3.2 — Анализ результатов разметки.

Категория	Кол-во синсетов
Соответствие найдено	86
Нет соответствия	9
Не найдено кандидатов	3
Агрессивная кластеризация	2

Таким образом, по результатам тестирования можно говорить о достаточно высокой эффективности предложенного метода автоматического выравнивания синсетов.

3.6. Неиспользованные подходы

Несколько потеницальных подходов к улучшению характеристик графа нами не были использованы. В настоящем разделе мы рассмотрим наиболее интересные из них и обсудим причины, почему они не были реализованы.

3.6.1. Анализ тезаурусных связей

Тезаурус — семантическая сеть, организующая лексику языка. Ключевое наблюдение в том, что это сеть, т.е. не просто множество синсетов, но, что важнее — связей между этими синсетами. Теоретически потенциал использования тезаурусных связей в задаче выравнивания огромен: если синсеты $a \in A$, $b \in B$ уже связаны, то, например, находя множество кандидатов для связывания синсета a' – гипонима a, мы можем ограничить область поиска гипонимами b. Это бы могло значительно улучшить эффективность автома-

тического выравнивания, т.к. полисемичнные слова относятся довольно часто к понятиям в разных участках графа, иными словами гиперонимы/холонимы существенно уточняют смысл слов синонимических рядов своих гипонимов/меронимов.

Как уже было замечено, на данный момент в YARN тезаурусные связи отсутствуют. Тестирование, однако, показало, что как минимум для синсетов из BCs это не является существенным затруднением. При попытке проведения в будущем выравнивания для более широкого множества понятий, при условии что к тому моменту тезаурусные связи будут представлены для значительного множества синсетов YARN, их использование представляется чрезвычайно перспективным шагом к улучшению результатов автоматического предпроцессинга.

3.6.2. Использование машинного перевода

На различных этапах разработки метрики нами рассматривалась возможность использования машинного перевода.

Первая идея его использования — для обработки определений, примеров использования. Т.е. при наличии определений в обоих синсетах можно перевести одно из них в язык другого, сравнить полученные строки как последовательности слов одного языка. В случае с примерами использования можно попробовать даже более интересный подход: так как в примерах встречаются слова синсета, погруженные в контекст использования, при удачном переводе полисемия будет должным образом разрешена (благодаря использованию контекста), и в случае, если в полученном переводе будет найдено одно из слов синсета кандидата, вес ребра между ним и исходным синсетом можно повысить (т.к. факт, что оно будет найдено в переводе — дополнительное свидетельство о связанности синсетов).

Эта идея не была нами реализованна в силу того, что полученная нами метрика и так показала достаточно хорошие результаты. Если задача повышения точности метрики вновь станет актуальной, эту идею стоит опробовать в действии.

Другой идеей применения машинного перевода является его использование в качестве дополнительного словаря. Некоторые синсеты PWN содержат не только слова, но многословные (обычно двусловные) устойчивые словосо-

четания. Таковые обыкновенно плохо представлены в переводных словарях, техники машинного перевода же прекрасно справляются с такого рода выражениями.

Однако, подключив сервис Яндекс.Словари, мы обнаружили, что все из рассмотренных нами подобного рода выражений содержатся в выдаче этого сервиса. Более того, структура ответа (отсутствие, как правило, примеров использования и единственный результат перевода) свидетельствует о том, что машинный перевод уже используется на уровне движка сервиса (сравнение выдачи с выдачей сервиса Яндекс.Перевод подтвердило эту догадку), и как следствие необходимость в подключении дополнительно сервиса машинного перевода отпала.

3.6.3. Использование статистических данных

Частотный словарь — словарь, в котором для каждого слова определяется некоторое число. Чем больше число, тем чаще данное слово встречается в языке (по факту - в некотором корпусе текстов,по которому составлялся данный словарь).

Авторы работы [16] предложили следующее использование частотного словаря в решении задачи автоматического выравнивания: из всех переводов слова (наборов синонимов) рассматривать в первую очередь те, максимальная частотность слова которых минимальна. Сделано это было из соображения, что менее частотные слова относятся к более частным синсетам, и как следствие четче показывают смысл исходного синсета.

Дистрибутивная семантика — область лингвистики, которая занимается вычислением степени семантической близости слов на основании их дистрибуционных признаков в больших массивах лингвистических данных. В частности, методы дистрибутивной семантики применяются в популярном инструменте word2Vec [2], возможность использования которого в построении метрики мы также рассматривали. Данная утилита принимает на вход корпус текстов, некоторый набор параметров (таких, как размерность выходного векторного пространства), строит отображение слов в вектора некоторого K-мерного пространства и по построенному пространству позволяет оценивать семантическую близость слов в предоставленном на вход корпусе.

Данные, предоставляемые этой утилитой, можно использовать в построении метрики. Например, при оценке близости множества слов перевода с множеством слов синсета (в дополнение к применяемой для этого мере Жаккара).

Не смотря на кажущуюся на первый взгляд эффективность использования методов, предоставляющих информацию о словах на основе статистики, мы не ожидаем значительного эффекта от их использования в построении метрики. Дело в том, что все эти методы имеют дело именно со словами (лексемами), никак не различая смыслы. Наибольшую же сложность в построении метрики представляет именно различение различных смыслов слова (разрешение полисемии). В силу этих соображений мы отказались от реализации использования этих методов, отдав предпочтение другим подходам.

3.7. Обработка BCs

Как было сказано в разделе 2.3, для тестирования процедуры выравнивания нами было выбрано множество синсетов, сформированных группой, работавшей над проектом BalkaNet, так называемых основных понятий (base concepts или BCs). Множество основных понятий формировалось как множество понятий, отчетливо присутствовавших во всех рассматривавшимися исследователями языках (болгарский, чешский, греческий, румынский, турецкий, сербский). Как следствие, можно предположить, что большинство из них будет присутствовать и в русском языке.

В общем случае для проведения выравнивания мы можем выбрать и произвольное множество синсетов PWN. Но тогда мы непременно столкнемся с значительно большими трудностями, чем при связывании основных синсетов. Основной, конечно, является фундаментальное различие лексических наборов двух языков: некоторые синсеты PWN описывают понятия, специфичные для английского языка и не представленные в четком виде в языке русском. При связывании синсетов из BCs при тестировании на 100 случайных синсетов PWN только для двух лингвист затруднился назвать соответствующее понятие русского языка. Для произвольного синсета PWN количество таковых может быть значительно больше (и кроме того, чем специфич-

нее синсет, тем с меньшей вероятностью соответвующее ему понятие русского языка будет представлено в YARN, даже если таковое существует).

Также усилятся проблемы, связанные с полисемией, неоднозначно определенными синсетами и неполнотой покрытия YARN, однако все эти трудности представляются преодолимыми. Чтобы не сталкиваться со всеми этими трудностями на первых этапах, мы решили выполнить связывание в первую очередь для BCs, а уже после предпринять попытку обобщения метода на более широкое множество синсетов.

Результаты проекта BalkaNet доступны в открытом доступе, в том числе доступно и выделенное множество основных понятий, в формате XML. Однако, поскольку сбор данных производился в 2003-2004 годах, понадобилась некоторая конвертация: идентификаторы синсетов PWN были даны относительно PWN версии 2.0. Актуальной же является версия 3.1, причём особенностью PWN является полное отсутствие обратной совместимости между идентификаторами разных версий. Потому, чтобы использовать BCs, нами была реализована вспомагательная утилита, которая на основании меры Жаккара сининоимических рядов и расстояния Левенштейна определений производит сопоставление синсетов между версиями PWN, а для тех, для которого такое сопоставление не удалось выполнить автоматически, позволяет выбрать соответствие вручную (из нескольких предложенных вариантов, всего вручную были обрабаотаны порядка 100 синсетов из 4500+ BCs).

Получив BCs в виде списка идентификаторов PWN 3.1, мы запустили для 3143 синсетов-существительных процедуру автоматического предпроцессинга. Распределение количества найденных кандидатов для различных синсетов дано в таблице 3.3.

Таблица 3.3 — Размеры множеств кандидатов для существительных из BCs.

Кол-во кандидатов	Кол-во синсетов	%
Всего	3143	100 %
0	194	6,2 %
$1 \leqslant 4$	579	18,4 %
$5 \leqslant 10$	572	18,2 %

Кол-во кандидатов	Кол-во синсетов	%
$11 \leqslant 15$	339	10,8 %
$\geqslant 16$	1459	46,4 %

Резюме

В данной главе был подробно рассмотрен предложенный нами метод алгоритмического решения задачи выравнивания. Было введено понятие графа связей — взвешенного ориентированного графа, построена метрика на нём, рассмотрены основные трудности, возникшие при её построении.

Проведено тестирование получившегося метода построения метрики, разобраны результаты. Рассмотрены подходы, неиспользуемые на данный момент, однако представляющие интерес для реализации (аппробации) в будущем. Выполнен процессинг 3143 существительных BCs с помощью получившегося метода, вкратце разобраны его результаты.

ГЛАВА 4. ПРИМЕНЕНИЕ КРАУДСОРСИНГА

Краудсорсинг — подход к получению необходимых сервисов, информации посредством объединения усилий широкого круга лиц, в особенности онлайн сообщества, в противовес традиционному использованию труда наёмных работников (специалистов). Сам термин происходит от английских crowd (толпа) и source (ресурс), что может быть интерпретировано, как получение ресурса силами “толпы”.

Краудсорсинг активно применяется в самых разных задачах, связанных с получением (обработкой) информации, в частности во многих приложениях области информационного поиска. Удобство развитых техник в том, что они позволяют снизить участие профессионалов в создании продукта до минимума, переложив большую часть работы на плечи сообщества. В различных приложениях лингвистики использование краудсорсинга представляет дополнительный интерес, предоставляя возможность получения данных “из первых рук”, непосредственно от носителей языка, позволяя ухватить интуитивное ощущение языка носителями (что может расходиться с ощущением языка профессионального лингвиста).

В данной работе методы краудсорсинга были использованы главным образом для разрешения полисемии. Как было замечено в разделе 3.3.2, в общем случае разрешить полисемию алгоритмически представляется достаточно трудной задачей. В то же время для любого носителя русского языка, знающего английский на базовом уровне, выбрать для данного синсета PWN соответствие из YARN представляется задачей несложной.

Далее в этой главе мы подробно рассмотрим применение краудсорсинга в предложенном нами методе выравнивания.

4.1. Предварительные требования

Рассмотрим основные требования, которые требуется удовлетворить, чтобы сделать возможным применение краудсорсинга к решению задачи.

Решение задачи выравнивания можно представить следующим образом: формируется пакет заданий, каждое соответствует синсету из PWN, некоторому множеству кандидатов. Респонденту предлагается выбрать соответ-

ствующие синсету PWN синсеты из YARN, либо отметить, что соответствие не найдено.

Во-первых очевидно следующее наблюдение: списки кандидатов должны быть приемлемого размера. Человеку будет сложно даже ознакомиться с тридцатью синсетами, тем более выбрать из них 1-3 наиболее подходящих.

Во-вторых, характерной особенностью краудсорсинга является анонимность участников, отсутствие достоверных данных об их квалификации (и существовании таковой вообще). Если б мы прибегли к помощи эксперта-лингвиста, мы с большой долей вероятности могли бы быть уверенными в достоверности данных им ответов. В случае с анонимным участником мы не можем быть уверенными ни в его квалификации, ни в его дисциплинированности при выполнении заданий.

Однако практика показывает, что если разбить задачи на множество малых и относительно простых подзадач, давать каждую задачу на исполнение нескольким людям (снижая возможный эффект ошибки респондента), а также проводить разметку в несколько этапов (ответ – валидация), можно получить ресурс достаточно высокого качества, сравнимого с качеством ресурса, полученного в результате привлечения квалифицированных исполнителей.

Во многом вследствие этих соображений при разработке процедуры автоматического выравнивания в качестве важнейшего критерия качества получаемого результата мы рассматривали размер получаемого множества синсетов. Множество размера 7–15 достаточно легко разбить на одно или несколько небольших заданий, выполнения каждого из которых займёт у респондента сравнительно небольшое количество времени.

4.2. Формулировка заданий

Задания на разметку мы сформулировали следующим образом: пользователю даётся подробная информация о синсете из PWN (слова, определение, примеры использования, изображения), предлагается выбрать ровно один наиболее близкий по смыслу синсет-кандидат из YARN, либо отметить, что таковых нет.

Вообще говоря, в YARN может содержаться более одного синсета, хорошо описывающего данное понятие английского языка (вследствие различия

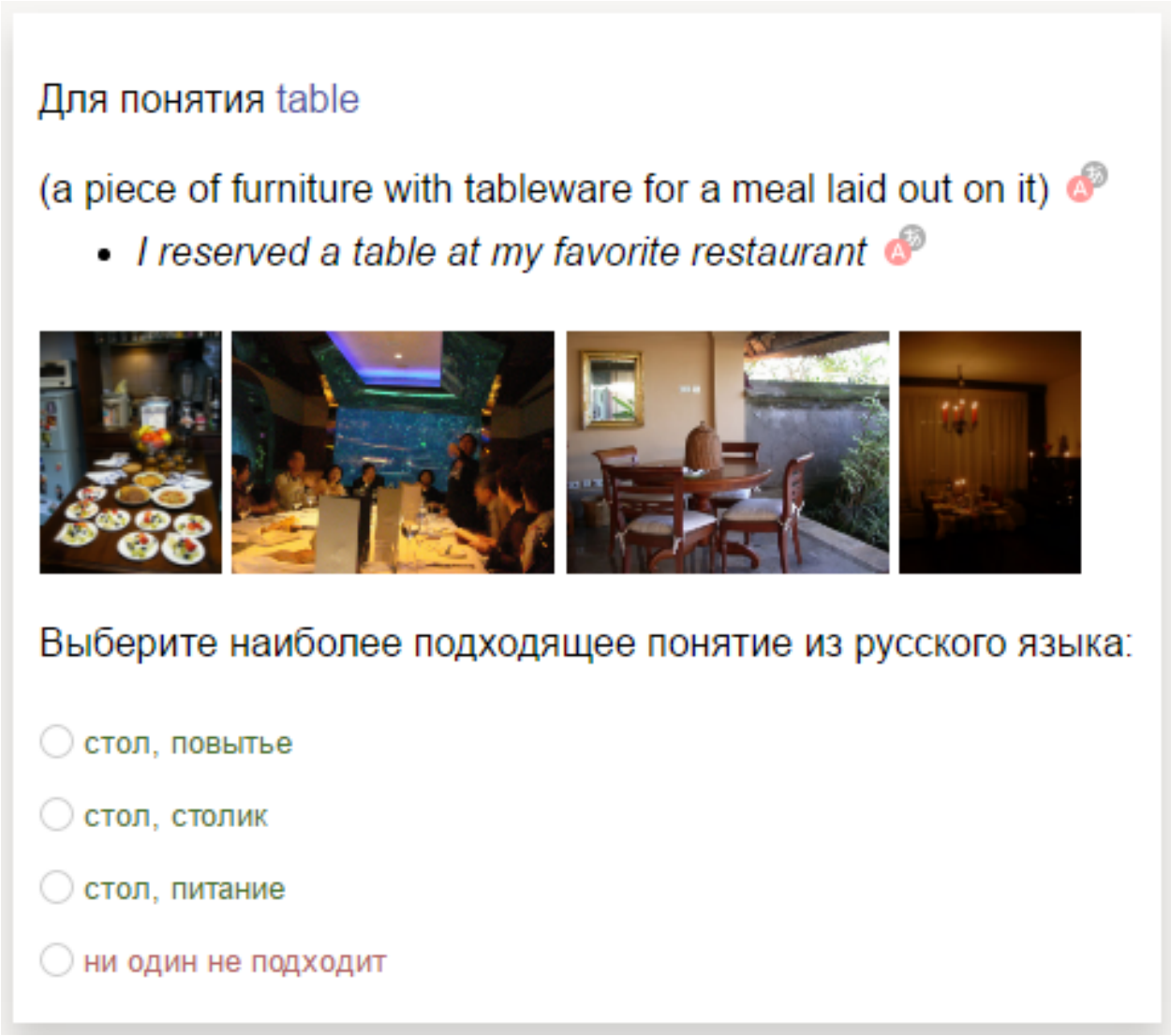

Рис. 4.1 — Интерфейс разметки пользователя в Яндекс.Толока

языков, дублирования синсетов), и естественной мыслью будет дать пользователю выбрать не один, а несколько наиболее подходящих кандидатов. Однако это привело бы к значительному усложнению заданий. Кроме того, некоторые пользователи были бы склонны отмечать все сколько-нибудь относящиеся к исходному понятию синсеты, другие же напротив — не более одного, что приводило бы к дополнительным трудностям в обработке данных. И поскольку поставленная нами задача допускает ослабление в виде выдачи ответа не как множества всех наиболее близких синсетов YARN, но как одного из элементов такого множества, формулировка заданий как однозначного выбора из множества кандидатов представляется вполне разумной.

На скриншоте 4.1 показан пример задания (в интерфейсе Яндекс.Толока). Описание, примеры использования взяты непосредственно из PWN, изображения взяты из проекта Imagenet [22]. Формируются наборы из не более чем

20 заданий, каждый набор показывается пользователю на отдельной странице, после выполнения всех заданий набора пользователь переходит к следующему. Среднее время выполнения набора по результатам тестирования составило 4 мин 30 сек, что приемлимо.

4.3. Рабочий цикл

Как было замечено в разделе 4.1, для формирования заданий мы сформулировали для каждого интересующего нас синсета PWN набор из не более чем 15 синсетов-кандидатов. Однако дать пользователю задание с пятнцадцатью вариантами ответов представляется идеей не самой разумной. Чем больше вариантов, тем как правило сложнее среди них определить правильный, и поэтому было решено включать в задание не более пяти вариантов ответа (плюс вариант “ни один не подходит”). Однако, как было замечено в разделе 3.5, для 7 синсетов из 100 (на которых проводилось тестирование) первый близкий кандидат был найден далее, чем на пятой позиции. Существует как минимум два варианта решения данного затруднения:

– ограничить число связей пятью — уменьшить покрытие метода (на 7%)
– рассматривать более пяти связей, разбивая полученное множество кандидатов на несколько заданий меньшего размера

Нами была выбрана вторая опция. Её выбор непосредственно влечет за собой увеличение общего числа заданий, однако как будет показано далее в этом разделе, увеличение происходит в допустимых пределах.

Кроме того, создавая для каждого синсета несколько заданий, у нас появляется необходимость разработать методологию для объединения результатов выполненных заданий, проведения дополнительных этапов тестирования. Такую методологию мы будем называть “рабочим циклом” или просто — циклом.

Итерацию цикла будем называть раундом. Процесс выравнивания заданного набора синсетов PWN будет состоять из нескольких раундов, на каждом из которых отсекается некоторое количество кандидатов (и при необходимости создаются задания для дополнительной фильтрации “выживших” кандидатов).

Пусть N_{max} — максимальное количество опций задания, E_{max} — максимальный размер исходного множества кандидатов. Для каждого синсета s из PWN к первому раунду генерируется $K_s = \frac{\min(E_s, E_{max})}{N_{max}}$ заданий. Множество кандидатов равномерно распределяется между этими K_s заданиями: i-й кандидат добавляется в j-ое задание, где $j = i \bmod K_s$.

После каждого раунда для каждого задания определяется не более W_{max} победителей. Из полученного множества победителей формируются новые задания (посредством объединения победителей различных заданий в новые задания, как кандидатов). Выравнивание считается завершенным, как только для синсета будет определено множество из не более чем W_{max} кандидатов-победителей, полученных из одного задания.

4.4. Обработка результатов

В предыдущем разделе было замечено, что для каждого задания по результатам тестирования определяется не более W_{max} победителей. Существует целое множеств методик агрегации результатов, т.е. позволяющих из полученных посредством краудсорсинга результатов определить варианты-ответа – “победители”.

Самым простым, очевидным методом является выбор победителя по большинству голосов. Т.е. если у нас есть задание с вариантами ответа $1, 2, 3, 4, 0$ с распределением голосов $0, 0.2, 0.4, 0, 0.4$, победителями будут выбраны варианты ответа 3, 0.

Однако помимо голосования большинства, существует целый ряд алгоритмов, решающих задачу агрегации, дающих на выходе значительно более релевантные результаты. В частности авторами бенчмарка SQUARE в работе [23] было проведено сравнение нескольких таких алгоритмов на нескольких публично доступных датасетах, во всех тестах алгоритмы давали результаты значительно лучшие полученных агрегацией методом большинства.

В настоящей работе было принято решение использовать фреймворк MTsar [24], в числе многих возможностей которого имеется простое для реализации API для агрегации результатов.

В MTsar представлены следующие алгоритмы агрегации результатов:

– KOS [25]

– Метод Давида-Скина [26]
– Zencrowd [27]
– Голосования большинства

Алгоритм KOS для решения нашей задачи не подходит, т.к. требует формулировки заданий в виде бинарного множества вариантов ответов. При проведении тестирования было проведено сравнение результатов, предоставляемыми остальными тремя методами агрегации. В результате наилучшие результаты были предоставлены методом Zencrowd. Метод Давида-Скина только для некоторых заданий давал более релерелвантные ответы, для большинства же заданий результаты агрегации более походили на шум, что могло свидетельствовать о расхождении метода на предоставляемых ему данных (наборы из 159 и 57 заданий с перекрытием 5).

Алгоритм Zencrowd на выходе даёт вектор вероятностей - оценку вероятностей правдоподобия каждого варианта ответа. Этот вектор представляет собой очень ценный источник информации. Например, определяя победителей, разумно было бы выбирать их только среди тех вариантов, в правдоподобии которых алгоритм “достаточно уверен”. Как следствие, вместо выбора W_{max} вариантов с максимальными вероятностями, нами было принято решение выбирать варианты с оценкой вероятности не менее $T = 0.4$ (из чего следует $W_{max} = 2$).

4.5. Тестирование

Как было изложено в разделе 3.5, для тестирования нами было выбрано случайное подмножество из 100 синсетов PWN, для 97 из которых было найдено хотя бы по одну кандидату на связывание. В этом разделе мы опишем ход и результаты второго этапа тестирования, исследующего эффективность процедуры выравнивания в целом (т.е. комбинацию автоматического выравнивания и применения краудсорсинга).

В процессе тестирования было проведено два раунда. В первом было обработано 159 заданий, во втором — 57 заданий. В обоих раундах было использовано перекрытие (количество раз, которое будет выполнено каждое задание разными рабочими) 5.

Задания для первого раунда были сформированы на основании графа связей. 16 из 57 заданий второго раунда также были получены только на основании связей, полученных автоматическим выравниванием (синсеты этих заданий не были включены в первый раунд из организационных соображений), другие 39 заданий второго раунда были сгенерированны на основании результатов первого раунда.

Как было отмечено в 3.5, тестирование метода проводилось путем сравнения с эталонной разметкой. С помощью краудсорсинга мы получили для каждого исходного синсета из PWN от 0 до 2 синсетов-победителей, сравнили полученные данные с эталоном. В таблице 4.1 представлено сравнение результатов.

Таблица 4.1 — Сравнение результатов краудсорсинга и экспертной разметки.

Результат	Кол-во
Нет перевода	3
Неверная классификация, “ни один не подходит”	1
Неверная классификация, нерелевантный синсет	12
Согласующаяся классификация, “ни один не подходит”	5
Согласующаяся классификация, релевантный синсет	79

Причем, из 79 классифицированных согласующимся с мнением эксперта образом синсетов, 56 получившихся классификаций были отмечены экспертом, как полностью соответствующие. Остальные 23 установленные связи требуют дополнительного уточнения синсетов из YARN, чтобы соответствующие синсеты также полностью соответствовали понятиям, описанным исходными синсетами PWN.

Однако установленные связи с синсетами, не полностью соответствующими понятиям (но достаточно близкими) — еще не столь большая проблема. То же касается и синсетов, для которых неверно помечено, что ни один кандидат не подходит. Гораздо более значительную проблему представляют синсеты, для которых были выбраны некорректные кандидаты. Мы подробно проанализировали причины неверной классификации:

– смешение понятий в синсете YARN: 1

– выбран нерелевантный синсет при отсутствии релевантного в выборке
  – выбран синсет, соответствующий английскому понятию, имеющим общие слова с исходным: 2
– выбран нерелевантный синсет при наличии релевантного в выборке
  – выбран синсет, соответствующий английскому понятию, имеющим общие слова с исходным: 5
  – выбран слишком общий гипероним: 1
  – выбран слишком частный гипоним: 3
– не выбрано ни одного при наличии релевантного синсета в выборке: 1

Что замечено, часто синсеты успешно проходили первый шаг (т.е. в множестве кандидатов по результатам голосования на первом шаге оставались в том числе релевантные синсеты), однако на втором участники голосования выбирали ошибочный синсет. На самом деле 12 синсетов (или 12% тестовой выборки) — достаточно неплохой результат. Однако для построения тезауруса (а задача выравнивания формулируется именно в как подзадача задачи построения тезауруса YARN) точность 88% не является удовлетворительной.

4.5.1. Сравнение с результатами других работ

Чтобы сравнить полученный результат с результатами, полученными авторами других работ, посчитаем значения точности (precision) и чувствительности (recall) по следующим формулам (found — найденные сопоставления, correct — корректные сопоставления):

$$\text{recall} = \frac{|\,\text{found} \cap \text{correct}\,|}{|\,\text{found}\,|}$$

$$\text{precision} = \frac{|\,\text{found} \cap \text{correct}\,|}{|\,\text{correct}\,|}$$

Получаем значения $\text{precision} = 0.91$, $\text{recall} = 0.86$.

Известные нам попытки решения задачи выравнивания тезаурусов различных языков были рассмотрены в разделе 2.2, все они производились силами экспертов. Существует множество работ, посвященных автоматическому выравниванию тезаурусов, онтологий созданных на одном языке. Проводились исследования и по автоматическому переводу тезаурусов в другие языки

(именно перевода тезауруса в другой язык, а не сопоставления двух независимо созданных тезаурусов).

Авторы работы [28] сравнивают эффективность различных методов в задаче соспоставления записей тезаурусов агрономической области. Заметим, что авторами рассматривается задача сопоставления тезаурусов для английского языка в узкой области применения (что значительно отличается от задачи, решаемой нами). Наилучшие значения точности, чувствительности, полученные в ходе их эксперимента: precision $= 0.84$, recall $= 0.49$.

Авторы работы [29] поставили своей задачей перевод тезауруса PWN на французский язык. Ими был применен подход с использованием машинного обучения, в котором определялись корректные пары (n, t), где n — синсет PWN, t — слово французского языка. Авторы измерили значения чувствительности, точности для разных частей речи и в частности для существительных ими были получены значения precision $= 0.84$, recall $= 0.88$.

Таким образом, можно судить о достаточной эффективности полученного метода в сравнении с результатами аналогичных работ. В следующем разделе мы рассмотрим подходы, применяя которые мы надеемся улучшить полученный результат (однако применение которых выходит за рамки настоящей работы).

4.6. Последующая работа

В предыдущем разделе были подробно разобраны результаты тестирования на случайно выделенном множестве из 100 BCs. Из 100 синсетов для 13 выравнивание было произведено ошибочным образом, причём для 12 был определен синсет, нерелевантный исходному понятию. Рассматривая эти синсеты, мы разбили их на несколько случаев, проанализировав которые мы пришли к некоторым соображениям, которые, будучи реализованными, помогут (по нашему мнению) существенно увеличить точность результата.

4.6.1. Уточнение синсетов-кандидатов

Как было отмечено в разделе 3.5, для двух синсетов PWN еще на этапе нахождения кандидатов были “потеряны” релевантные синсеты YARN как

результат слишком агрессивного поведения эвристики объединения в кластеры. Работа над методами устранения дубликации в YARN сейчас активно ведется. Когда проблема дубликации будет решена, необходимость в эвристике объединения кандидатов в кластеры отпадёт.

Кроме того, в одном случае неверная классификация была вызвана наличием в списке кандидатов смешанного синсета (т.е. содержащего слова из нескольких понятий языка). Работа над YARN продолжается, и со временем доля подобных синсетов будет сведена к минимуму.

Некоторые ошибки были допущены между синсетами, связанными связями гиперонимии. После интеграции в YARN родо-видовых связей, применение которых для улучшения этапа автоматического выравнивания было рассмотрено в разделе 3.6, их также можно будет применить для обнаружения ситуаций, когда выбран слишком общий гипероним/слишком частный гипоним.

4.6.2. Улучшение процедуры краудсорсинга

Однако большинство ошибок было произведено не вследствие недостаточной чистоты/покрытия YARN, но непосредственно вследствие ошибок респондентов при классификации синсетов. Для 7 из 12 был выбран синсет, соответствующий английскому понятию, имеющему с исходным синсетом общие слова. Обращая внимание на все слова синсета и определение, можно сделать однозначный выбор в пользу другого кандидата, однако выбор был сделан ошибочно, по нашему наблюдению именно вследствие концентрации внимания исполнителя на многозначных словах синсета, невосприятия исполнителем информации о синсете в целом.

Простой попыткой решения проблемы может явиться модификация параметров проведения краудсорсинга:

– увеличение перекрытия
– уменьшение числа опций в задании
– увеличение оплаты за задание, уменьшение числа заданий в наборе

Кроме того, можно пересмотреть инструкцию, добавить обучающие примеры, ввести поощрения для пользователей, показывающих наилучшую результативность (соответствующее ранжирование предоставляется большинством методов агрегирования результатов).

Однако значительно более перспективным направлением исследования является разработка дополнительных видов заданий. В работах [30], [17] авторы предлагают многоступенчатые циклы проведения краудсорсинга, включающие этапы как сбора информации, так и валидации полученных данных. Нам кажется разумным пойти тем же путём.

В частности в ближайшее время планируется ввести и протестировать эффективность следующего вида заданий: показывать пользователю только определение синсета из PWN и одного или нескольких связанных синсетов из YARN. Пользователю может быть предложено либо подтвердить, что предложенный синсет соответствует определению, либо опровергнуть. Данная задача будет несколько сложнее задачи, сформулированной в разделе 4.2 тем, что исполнитель будет лишен возможности сориентироваться по предложенным словам. Однако тот факт, что вместо 3–5 синсетов-кандидатов на выбор будет представлен только один, напротив значительно упрощает задание. Отсутствие какой-либо информации, кроме определений (и, может быть, примеров использования), заставит пользователя разобраться в значении синсета, полисемия слов синсета из PWN в таком случае не внесет путаницы.

Также некоторая вариативность присутствует и в, собственно, характере ответа. Chris Bienmann в своей работе [30], например, предложил метод обнаружения лексических подстановок с использованием краудсорсинга, задания в котором формулировались не в терминах выбора наилучшего ответа, но в терминах выбора наихудшего (и таким образом, постепенно отсекались нерелевантные ответы).

В решении задачи выравнивания подобный подход показался нам менее эффективным, поскольку как правило в множестве кандидатов у нас располагается 1–3 релевантных синсетов и произвольное количество синсетов, представляющих другие понятия языка. Однако интерес представляет идея комбинирования заданий с выбором наилучшего и заданий с выбором наихудшего, например в первом раунде использовать голосование с выбором наилучшего, а на втором, когда остаётся менее 5 кандидатов — голосование с выбором наихудшего (возможно, проведенное в несколько итераций). Это наверняка увеличит количество заданий, однако возможно позволит существенно увеличить точность.

4.6.3. Интеграция связей в YARN

Однако, навряд ли хоть один подход (кроме проверки данных экспертом) сможет гарантировать нам точность, близкую к 100%. Потому всегда в нашем распоряжении будет какое-то количество связей, в действительности которых мы не сможем быть уверенными до конца, и как следствие следует оставить возможность их изменения в будущем.

Редактирование синсетов YARN проходит с использованием специально разработанного для этих целей интерфейса. Этот интерфейс возможно дополнить отображением и возможностью редактирования связанного понятия PWN. Некоторые связи, о достоверности которых мы не моем судить, могут быть помечены, как неподтвержденные. В дальнейшем редактор сможет либо подтвердить их, либо удалить или заменить на связи с другими синсетами из PWN.

Резюме

В настоящей главе было подробно рассмотрено применение краудсорсинга для решения задачи выравнивания (как второго этапа после автоматического предпроцессинга). Был сформулирован подход к разбиению задачи выравнивания на небольшие задания, которые вследствие их простоты можно использовать для краудсорсинга, сформулирован формат заданий, разработан интерфейс. Были рассмотрены алгоритмы агрегации результатов, проведено их сравнение.

Было проведено полномасштабное тестирование полученной процедуры выравнивания (обоих этапов), получены удовлетворительно высокие результаты. Наконец, были рассмотрены подходы, благодаря которым мы в будущем ожидаем повысить точность полученного результата.

ЗАКЛЮЧЕНИЕ

В настоящей работе была рассмотрена задача выравнивания синсетов тезаурусов Princeton WordNet и YARN. Был предложен метод для решения этой задачи, состоящий из двух этапов: автоматический предпроцессинг и выравнивание с применением техник краудсорсинга. Каждый из этапов был подробно рассмотрен в главах 3, 4.

Полученный метод был протестирован на случайном подмножестве существительных из BCs, были получены хорошие результаты как для автоматического выравнивания (см. раздел 3.5), так и для выравнивания с применением краудсорсинга. Итоговая точность метода на тестовой выборке составила 88%. Были рассмотрены направления для дальнейших исследований (нацеленных на повышение точности).

После достижения достаточной точности порядка 95% (как с помощью применения подходов, рассмотренных в 4.6, так и других), планируется провести процедуру выравнивнивания для всех существительных BCs, интегрировать полученные связи в структуру тезауруса YARN.

В дальнейшем планируется адаптировать полученный в настоящей работе метод для проведения выравнивания по произвольному множеству синсетов PWN.

СПИСОК ИСТОЧНИКОВ

1. Mystem. [Электронный ресурс]. URL: https://tech.yandex.ru/mystem/.
2. Word2Vec. [Электронный ресурс]. URL: https://code.google.com/archive/p/word2vec/.
3. OpenCorpora. [Электронный ресурс]. URL: http://opencorpora.org/.
4. Лукашевич Наталья Валентинова. Тезаурусы в задачах информационного поиска. М.: Издательство МГУ, 2011, 2010.
5. Fellbaum Christiane. WordNet. Wiley Online Library, 1998.
6. Miller George A. Nouns in WordNet: a lexical inheritance system // International journal of Lexicography. 1990. Т. 3, № 4. С. 245–264.
7. Fellbaum Christiane, Gross Derek, Miller Katherine. Adjectives in wordnet. 1993.
8. Fellbaum Christiane. English verbs as a semantic net // International Journal of Lexicography. 1990. Т. 3, № 4. С. 278–301.
9. Russnet: Building a lexical database for the russian language / Irina Azarova, Olga Mitrofanova, Anna Sinopalnikova [и др.] // Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas. 2002. С. 60–64.
10. Рутез: о проекте. [Электронный ресурс]. URL: http://www.labinform.ru/pub/ruthes/about.htm.
11. Тезаурус русского языка в формате WordNet - RuWordNet. [Электронный ресурс]. URL: http://www.labinform.ru/pub/ruwordnet/index.htm.
12. Braslavski Pavel, Ustalov Dmitry, Mukhin Mikhail. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus // EACL 2014 / Association for Computational Linguistics. 2014. С. 101–104.
13. Vossen Piek. EuroWordnet general document (Version 3–Final) // University of Amsterdam. EuroWordNet LE2-4003, LE4-8328. 1999.
14. The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology / Horacio Rodríguez, Salvador Climent,

Piek Vossen [и др.] // Computers and the Humanities. 1998. Т. 32, № 2-3. С. 117–152.

15. Română Bukarest Academia. Romanian Journal on Information Science and Technology, Special Issue on BalkaNet. Publishing House of the Romanian Acad., 2004. Т. 7.
16. Автоматический перевод семантической сети WordNet на русский язык / ИГ Гельфейнбейн, АВ Гончарук, ВП Лехельт [и др.] // Труды Международного семинара Диалог по компьютерной лингвистике и её приложениям, Протвино, Россия. 2003.
17. Ustalov Dmitry, Kiselev Yuri. Add-Remove-Confirm: Crowdsourcing Synset Cleansing // Application of Information and Communication Technologies (AICT), 2015 9th International Conference on / IEEE. 2015. С. 143–147.
18. Russian Lexicographic Landscape: a Tale of 12 Dictionaries / Yu Kiselev, A Krizhanovsky, P Braslavski [и др.]. 2015.
19. The Romanian Wordnet / Dan Tufi, Eduard Barbu, Verginica Barbu Mititelu [и др.] // Romanian Journal on Information Science and Technology, Special Issue on BalkaNet. 2004. Т. 7. С. 107–124.
20. Словари в формате StarDict. [Электронный ресурс]. URL: `http://download.huzheng.org/lingvo/`.
21. Krizhanovsky AA. Transformation of Wiktionary entry structure into tables and relations in a relational database schema // arXiv preprint arXiv:1011.1368. 2010.
22. Imagenet: A large-scale hierarchical image database / Jia Deng, Wei Dong, Richard Socher [и др.] // Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on / IEEE. 2009. С. 248–255.
23. Sheshadri Aashish, Lease Matthew. SQUARE: A benchmark for research on computing crowd consensus // First AAAI Conference on Human Computation and Crowdsourcing. 2013.
24. Ustalov D. A Crowdsourcing Engine for Mechanized Labor // Proceedings of the Institute for System Programming. Moscow, 2015. Vol. 27, no. 3. P. 351–364. URL: `http://www.ispras.ru/proceedings/docs/2015/27/3/isp_27_2015_3_351.pdf`.
25. Karger David R, Oh Sewoong, Shah Devavrat. Iterative learning for reliable crowdsourcing systems // Advances in neural information processing systems.

2011. С. 1953–1961.
26. Dawid Alexander Philip, Skene Allan M. Maximum likelihood estimation of observer error-rates using the EM algorithm // Applied statistics. 1979. С. 20–28.
27. Demartini Gianluca, Difallah Djellel Eddine, Cudré-Mauroux Philippe. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking // Proceedings of the 21st international conference on World Wide Web / ACM. 2012. С. 469–478.
28. Comparing human and automatic thesaurus mapping approaches in the agricultural domain / Boris Lauser, Gudrun Johannsen, Caterina Caracciolo [и др.] // Universitätsverlag Göttingen. 2008. с. 43.
29. De Melo Gerard, Weikum Gerhard. Mapping Roget's Thesaurus and WordNet to French. // LREC / Citeseer. 2008.
30. Biemann Chris. Creating a system for lexical substitutions from scratch using crowdsourcing // Language Resources and Evaluation. 2013. Т. 47, № 1. С. 97–122.