Newsblaster Russian-English Clustering Performance Analysis

Lawrence J. Leftin
Department of Computer Science
Columbia University, New York

Abstract: *The Natural Language Group is developing a multi-language version of Columbia Newsblaster, a program that generates summaries of news articles collected from web sites. Newsblaster currently processes articles in Arabic, Japanese, Portuguese, Spanish, and Russian, as well as English. This report outlines the Russian language processing software, focusing on machine translation and document clustering. Russian-English clustering results are analyzed and indicate encouraging inter-language and intra-language performance.*

INTRODUCTION

The Natural Language Group is developing a multi-language version of Columbia Newsblaster [1], a utility that generates summaries of news articles collected from web sites. The system facilitates online news browsing by topically clustering the articles and providing links to the originals. Consequently, a very large number of news stories from diverse sites can be reviewed in a short time.

Newsblaster currently processes articles in Arabic, Japanese, Portuguese, Spanish, and Russian, as well as English. This report outlines the Russian language processing software, focusing on machine translation and document clustering. Russian-English clustering results are analyzed and indicate encouraging inter-language and intra-language performance.

Newsblaster processing comprises the following phases:

- crawling,
- article extraction,
- clustering,
- summarization,
- classification, and
- web page generation.

During the crawling phase, Newsblaster collects web page contents from news sites visited and stores the information in a 'zeep' database.

In the article extraction phase, Newsblaster discriminates real articles from unwanted verbiage using a machine learning algorithm. Unwanted information comprises HTML directives, lists of links, and advertisements.

The clustering phase endeavors to group together articles that are topically related. The articles are grouped into 'superclusters' and 'subclusters' for hierarchical access.

During the summarization phase, automatic textual summaries are produced for each cluster and article.

In the classification phase clusters and articles are classified and grouped into specified categories.

The web page generation phase organizes and displays this hierarchically organized information in an easily browsable format.

Multilingual Newsblaster

Previously, Newsblaster only processed articles written in English and encoded in ISO-8859-1 or UTF encodings. In addition, its clustering and summarization software processed the articles assuming English language grammar and word distributions. Consequently, a number of changes were necessary for accommodating Russian language articles. These changes included:

- accommodating Cyrillic encodings (windows-1251, KOI-18-R, and KOI-18U)
- training ArticleExtractor to discriminate real Russian articles
- translating Russian articles into English

The requisite Cyrillic encoding handlers were incorporated and allowed Newsblaster to collect and extract articles from virtually any Russian web site. The author then trained Newsblaster's Article Extraction module on a corpus of 112 Russian news articles with a precision of 90.588% and a recall of 95.06%. Both the encoding changes and article extraction performance are described in Klavans and Evans [2]. The current report focuses on the Russian translation- and Russian-English clustering aspects of the new system.

The next section provides a brief description of the challenges involved in multilingual clustering and efficient machine translation of Russian articles. It also provides an introduction to the Russian language, its relation to other Indo-European languages, its grammar and structure. The study then discusses current Russian-English clustering performance including initial assumptions, evaluation criteria, and detailed procedure. The Results section presents detailed (cluster and supercluster) performance as well as summary statistics. We end this report with a discussion of the experimental results and suggestions for future research.

BACKGROUND

Newsblaster's document clustering system (Hatzivassiloglou et al., 2000) employs a machine learning algorithm and has been trained on a large corpus of English articles. Rather than train the system in other languages, our approach is to translate the non-English language articles into English, then cluster all articles using the current clustering system, already trained in English. This allows multilingual document clustering without an additional inter-language clustering step.

While a highly accurate machine translation might seem the best approach, experience has shown this to be excessively slow for the volume of articles crawled. Instead, a simpler, dictionary-based Russian translation system is being employed. Because Russian is a highly inflected language, with a multiplicity of forms for cases and tenses, most words in an article will not automatically match those in a given glossary. In addition, there is appreciable 'intersection' amongst word forms arising from the inflections. Consequently, considerable effort was taken to match a given word form to the most probable glossary translation. If the simple glossing system proves adequate as a pre-clustering step, only those articles actually clustered will be precisely translated.

The remainder of this section provides an introduction to the Russian language, the design of the Russian machine translator, and a description of clustering experiments performed.

Russian Language

Russian, a member of the Slavic family of Indo-European group, is the native language of 167,000,000 people world-wide [4]. It is closely related to other Slavic languages including the West Slavic languages Polish, Czech, and Slovak, the East Slavic languages including Ukrainian and Byelorussian, and the South Slavic languages Serbian, Croatian, Slovenian, and Bulgarian. Linguists regard the Slavic family closely related to the Baltic family (Latvian and Lithuanian) but do not agree on whether this is due to a common Proto-Slavic-Baltic origin or simply due to sharing and close geographic proximity [5]. Likewise, the Slavic family bears a closer relationship to the Germanic family than to the other Indo-European families such as the Italic or Indo-Iranian families [5]. A rule of thumb for identifying Indo-European languages is testing whether their words for numbers, the verb "to be", or words for family members closely resemble equivalent words in other Indo-European languages. References [5] and [13] reveal Russian's obvious relation to other Indo-European languages via words for numbers and forms of the verb "to be", respectively.

Russian, like its Slavic sisters, is highly inflected and exhibits distinct word forms for nouns, pronouns, and adjectives according to case, gender, and number and distinct forms for person, number, tense, aspect, and mood of verbs. Russian nouns, pronouns, and adjectives fall into six cases as follows:

- nominative (subjects of sentences, predicate nominatives)
- genitive (indicates possession, close relationship, object of certain prepositions and verbs)
- dative (indirect objects, objects of certain prepositions)
- accusative (direct objects, objects of certain prepositions)
- instrumental (indicates instrument or agent, object of certain prepositions)
- locative (object of certain prepositions, especially indicating 'place at which')

Russian nouns occasionally exhibit a secondary genitive (partitive genitive) and a secondary locative. Russian verbs provide a full complement of forms for present-, past-, and future tenses, for both imperfective- and perfective aspects. Unlike English, Russian aspect is embodied in special verb forms and does not require helping verbs such as "have" or "has". Russian is also characterized by its essential lack of articles, which is in sharp contrast to Spanish and Portuguese or even English.

Russian is a rather musical language and is replete with single-vowel endings for noun cases and double-vowel endings for adjectives. This feature together with the various other inflections give Russian its characteristic rhythm and reaches a pinnacle of beauty in Russian poetry of Pushkin, and the prosody of Dostoevsky, Tolstoy, Turgenev, and Gogol.

Russian Translation System

The comparative richness of word forms also complicates machine translation. Consequently, the probability of a given word form matching the 'base' form is much smaller than in English, which has almost no inflection to speak of. We therefore needed an approach to deal with Russian inflections, while not going to the extreme of a full-blown translation.

A 'dictionary based' approach was taken, modified to mitigate the multi-wordform problem. In fact, the system uses five dictionaries/glossaries as follows:

1. 1800 Most Frequent Russian Word Forms (Inflected Forms)
2. 5000 Most Frequent Russian Words (Base Forms)
3. 398 Russian Place Names Glossary (Base Forms)
4. 988 Russian Personal Names Glossary (Base Forms)
5. 150,000 Entry Russian Dictionary (Base Forms)

The first dictionary [7] should match the majority of written Russian words exactly without the need for stemming. Consequently, this dictionary is consulted first for efficiency. Since many of the most frequent words are small, this prevents them from confounding stem-based matching logic that follows.

The second dictionary [8] is meant to catch significantly more commonly written words. Since part of speech information was present, the dictionary entries were stemmed ahead of time. A candidate word then matches the entry if it starts with the stem.

The Russian Place Names Glossary [9] is a compendium of about 400 place names gleaned from web searches. Place names were singled out because they are important article subject discriminants.

The Russian Personal Names Glossary [10], [11] was likewise compiled from web-based searches and provides both native Russian names and Russian equivalents of non-Russian names. Personal names are important article subject discriminants

The large Russian dictionary [12] is very comprehensive and typically catches any words missed by the first four dictionaries. It is placed last for efficiency.

The dictionaries are pre-sorted and loaded into memory at program initialization. Since the large dictionary had no POS information, no attempt was made to POS tag the input articles.

The program loops through every article word performing the following checks:

- is the word is a number or numerical date?
- does the word match one of the 1800 most frequent Russian word forms (these include exact inflections)?
- does the word match one of the 5000 most frequent Russian words (base form)?
- is the word a Place Name?
- is the word one of the most frequent Personal Names?
- does the word match one of the 150,000 words in the large Russian dictionary?
- if the word matches no dictionary entry, transliterate it.

For dictionaries 3, 4, and 5 we needed to perform 'naive stemming', in order to match the various possible inflected word forms. This was done as follows:

- try to match the current word form exactly to the dictionary word
- try to match n-1 characters of the current word to the dictionary word
- try to match n-2 characters of the current word to the dictionary word
- try to match n-3 characters of the current word to the dictionary word
- try to match the current word to the first m-1 characters of the dictionary word
- try to match n-1 characters of the current word to to the first m-1 characters of the dictionary word
- try to match n-2 characters of the current word to the first m-1 characters of the dictionary word
- try to match n-3 characters of the current word to the first m-1 characters of the dictionary words

where n = length of the current article word
m = length of the dictionary word

The above algorithm matches the idiosyncracies of the Russian inflections as follows:

- Most singular nouns have 1-character endings (except 0 for masc. nominative, 2 for instrumentals)
- Most plural nouns have 2 -character endings (except 1 for nominative, 3 for instrumental)
- Most adjectives have 2-character

endings (except 3 for masc/neut. genitive and dative singular, 3 for instrumental pl.)
- Most present/perfective verbs have double endings (except singular and 2nd pers. plural).

ARTICLE CLUSTERING

As noted in the previous section we have elected to translate foreign articles into English, then run the current clustering system on the resultant article set. This makes best use of the extensive training /testing already done and eliminates the need for an extra inter-language clustering step. This approach must be experimentally verified using real datasets. Consequently, a series of experiments have been designed and run for that purpose.

A full description of the experimental approach, procedure, and results appears in the following section.

EXPERIMENTAL APPROACH AND RESULTS

A series of experiments were run to determine clustering performance over a multilingual corpus of 880 articles. The input consisted of 226 English language articles, and 554 Russian language articles. The articles were obtained by running the (multilingual upgraded) Crawler and ArticleExtractor components of Newsblaster over the sites www.cnn.com, www.izvestia.com, and www.mn.com on March 25, 2003.

In experiments 1-4, and 6, the Russian articles were word-for-word translated using the glossing system described above. In experiment 5, the articles were translated using SYSTRAN's full-featured translation system (access provided by Babelfish).

After performing experiments 1-4, the author inspected the translation output and tried to identify the failure mode(s) in the clustering performance using the Russian word-for-word translator It turned out that a 4% error rate was traceable to two factors:

- 'small' words miscorrelating with larger ones
- Proper names not in the database mistranslated to other words

(Of course other inadequacies contribute to a higher overall error rate, but some of those have little effect on clustering.)

All the mistranslated words were red-lined, then located them in the original Russian. Often the same words popped up again and again. Next, the mistranslated words added either to the 'simple Russian wordform list', whose words are checked first for an exact match or to the 'Russian Place Names List' or 'Russian Names List', whose words are checked next. This approach intercepts the red herrings by translating them correctly before they can miscorrelate in the naive stemming logic that follows, and effectively 'trains' the glossing system.

The articles were clustered by executing the current clustering software on the resultant article set. The clustering program organizes the articles into superclusters and subclusters, which are listed in the output file, cluster_output.txt. The author then evaluated the quality of the output clusters by reading the articles in their respective source languages and manually categorizing them.

Newsblaster team members previously noted that articles from the same language tended to cluster more than articles from different languages. We therefore attempted to pre-compensate for the 'translation barrier' by amplifying the most salient words from the articles (e.g. proper nouns). Since proper nouns such as place names and personal names require little or no word sense disambiguation, the Russian word-for-word translator should translate these items best.

The following clustering tests were performed to determine performance over a variety of settings:

1. Original Article Set
2. Russian proper nouns repeated twice, CNN articles unchanged
3. Russian proper nouns repeated thrice, CNN articles unchanged
4. Russian- and CNN proper nouns repeated twice
5. Russian Articles Translated by Babelfish
6. Articles Translated by Newly-Trained Word-for-Word Translator

Experiment 1 - Untrained Baseline

The baseline (non-amplified) clustering results were inspected in detail. In nearly all cases the multi-language clusters collected articles that were very closely related.

Experiments 2-4 - Salience Amplification

In all three amplified cases clustering performance was considerably worse than the unadulterated (baseline) run. The negative results may be, in part, due to the large (x2, and x3) amplification factors implemented. It may turn out that more subtle adjustment of weights could yield a better result.

Experiment 5 - Babelfish Translation

In Experiment 5 the Babelfish online translation system that is based on SYSTRAN was used to translate the Russian articles. The Babelfish translation system provides a full-featured article translation complete with word-sense disambiguation and proper grammatical constructs. Clustering results using the Babelfish translator should provide a standard for evaluating the dictionary-based approach.

After the original Russian articles (from the March 25, 2003 corpus) were translated using Babelfish, the author ran clustering over the resultant set. Every article in every multilingual cluster was inspected in detail and evaluated for quality and quantity.

Experiment 6 - Trained Baseline

As described above, manual analysis of word-for-word translation errors provided feedback for upgrading the input dictionaries. The resulting article set was clustered and showed significantly better performance than before, albeit not as good as the Babelfish approach.

Results Summary

The results of the tests appear in Appendix A, *Experimental Results*. Each table summarizes the clustering results for each experiment. Each cluster is characterized by its supercluster ID, the number of English (CNN) articles in the cluster, the number of Russian articles in the cluster, a description of the subjects/topics appearing in the articles, and a 'grade' for the quality of the supercluster.

Results using Babelfish translation appear in Table A-5 wherein clusters identified by both systems are highlighted in red, while monolingual subclusters identical for each system are highlighted in green.

Results indicate:

1. 20 multilingual clusters identified when using Babelfish vs. 7 using the untrained dictionary-based translator and 12 for the trained dictionary-based translator
2. 79 CNN and 68 Russian articles appear in multilingual clusters using Babelfish vs. 24 CNN and 31 Russian articles using the untrained dictionary- based translator and 43 CNN and 43 Russian articles using the trained dictionary-based translator
3. Mono-lingual clustering performance is almost identical for both translation approaches.
4. The clusters identified using the dictionary-based approach are of high quality as are those using Babelfish.
5. The problem with the dictionary-based performance is one of omission of multilingual clusters, not of producing 'bad clusters'.
6. Using the trained dictionary-based translator, clustering performance was significantly better than before, albeit not as good as the Babelfish approach.

A top-level summary of experimental results appears in Table I. This table lists each experiment by number, the total number of multilingual clusters identified, and the percentage of good clusters produced.

The monolingual clustering performance validates D. Evan’s original premise that word-for-word translation is quite adequate for clustering articles. The degradation of multilingual performance was indeed partially alleviated by expanding the dictionaries of proper nouns and by moving previously miscorrelated words into the first dictionary. It is not expected that any dictionary-based approach would fully match Babelfish-based performance, though.

Table I. Summary Results - Multilingual Clustering Experiments

Experiment #	Description	# Multilingual Clusters	% Good Clusters
1	Untrained Word-for-Word Translation	7	100
2	Russian Words Repeated Twice	9	67
3	Russian Words Repeated Thrice	10	27
4	Russian and English Words Repeated Twice	14	33
5	Babelfish Translation	20	85
6	Trained Word-for-Word Translation	12	92

DISCUSSION AND SUGGESTIONS FOR FURTHER STUDY

Newsblaster Clustering Performance

Early clustering experiments confirm that Newsblaster can successfully cluster articles from diverse languages, provided the foreign articles are first translated into English. Detailed inspection of each multilingual Russian/English cluster indicate the member documents are very closely related. Monolingual clustering performance was excellent as well. We observed that the great majority of clusters were monolingual. This is, in part, because most local- and regional news stories pertain to the host country. It may also indicate a 'translation barrier' caused by the simplistic word-for-word translation method employed.

Clustering performance using a commercial-grade translation system was clearly superior to that using a dictionary-based translation. The Babelfish translation, however, took 2 hours to complete vs. 15 minutes for the dictionary-based system. The 8X speed improvement indicates the dictionary-based system would yield ~5 times the number of multilingual clusters for a given time span, even though a number of high-quality clusters would be missed.

Linguistic Radars and Document Interferometry

There is a parallelism between the world of document queries and that of radars. Radar is a form of electromagnetic echo location wherein an RF pulse is emitted, impinges on material objects, and is reflected back to the radar receiver. The returned signal strength is inversely proportional to the distance (to the fourth power) to the target and directly proportional to the transmitted power and to the target's radar cross section. A document query system emits a document query (pulse), which impinges on documents (targets), and returns documents whose degree of matching (signal strength) exceeds a threshold.

Unlike current query systems, reflected radar pulses are generally out of phase with the transmitted pulse. Pulse detection is done by correlating the input signal with a replica and digitizing the result. If only one replica were used, the detected signal strength would be a function of the relative phase. Consequently, real radar systems employ two pulse replicas (I and Q), which are 90 degrees out of phase. Portions of the signal not recovered by one replica will be retrieved by correlation with the other.

Surprisingly, words too, exhibit phase - A typical language has on the order of 100,000 words. While subsets of these words are conceptually distinct, many words differ only in shades of meaning or degree. We can therefore represent a given word by a base word and a percentage of the 'distance' to its antonym. For example consider the sequence:

exuberant joyous happy indifferent sad dejected devastated

Each member of the set could be represented as (exuberant, phase) where phase = 0.0 for "exuberant" and 180 for "devastated". Of course not all word sets are bipolar, so the ultimate representation for many word groups would employ di-hedral or even multihedral phases. The human mind cannot generally deal with more than seven objects at a time - consequently, there should be few word groups having more than 3 or 4 poles.

Employing a word-base-vector/phase model could vastly reduce the size of the 'word bags' used for document clustering (and queries) and hence improve performance. It also paves the way for 'coherent document processing', including looking at the constructive and destructive superposition of document pairs and document sets.

REFERENCES

[1] K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman, S. Teufel, (2001) Columbia Multi-document Summarization: Approach and Evaluation, Columbia University

[2] David Kirk Evans and Judith L. Klavans (2003_ A Platform for Multilingual News Summarization, Columbia University

[3] N. Wacholder, D. Evans, J. Klavans (2001) Automatic Identification and Organization of Index Terms for Interactive Browsing , Columbia University

[4] Summer Institute of Linguistics, Ethnologue - Languages of the World Web Pages, SIL International, 2003

[5] J.P. Mallory, In Search of the Indo-Europeans - Language, Archeology, Myth, Thames and Hudson, London, 1989

[6] Daniel Jurafsky, Martin, James H., (2001) Speech and Language Processing, Prentice Hall, Upper Saddle River, NJ

[7] Sergei Sharoff, Russian Frequency Lists: (web pages), http://www.artint.ru/projects/frqlist/frqlist-en.asp

[8] Superliminal Software, Dictionary of 5000 Most Frequent Russian Words, http://www.superliminal.com/tutor/tutor.htm

[9] Space Monitoring Information Support laboratory (SMIS IKI RAN), Russian Weather Pages

[10] http://slovo.and.ru/n-ya.htm, Russian First Names (web pages)

[11] http://www.bestseller.pp.ru/top_autor.php, American Russianized Names (web pages)

[12] www.slovnyk.org, Large Russian Dictionary - ru_RU-en_US.dict.koi8u.bz2, http://www.slovnyk.org.ua/prg/gszotar/koi8u/

[13] http://www.zompist.com/euro.htm, Table of Indo-European Numbers

[14]http://faculty.ccc.edu/trwebprojects/anushevskiy/assignment02/index.asp, Alexander Segeivich Pushkin Page

[15] Andreas Teuber, Dostoevsky Page,

http://people.brandeis.edu/~teuber/dostoevskybio.html#BiographicalInfoEssay

[16] J. Lyman, Tolstoy Page, http://www.ltolstoy.com/

[17] Eric Eldred, Turgenev Page, http://209.11.144.65/eldritchpress/ist/turgenev.htm

[18] Lazlo Tikos, Gogol Page, http://www.samizdat.com/gogol.html

APPENDIX A:

Experimental Results

Table A-1. English-Russian Clustering Performance - Baseline (Unmodified) Articles

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
12	4	5	Academy Awards, Scorcese's Oscar Prospects, Others Oscar Prospects, Oscars during war in Iraq	A
14	6	4	War in Iraq, Northern Iraq, Turkish reservations about Kurds, Stategic/Economic importance of Oil in North Iraq/Kirkut, Kurdish views on idependence, Ansar-Al-Islam-Al Queda link	A
27	3	4	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
47	1	14	CNN - senate/congress budget deliberations including tax cut; Russian: tax on automobiles and other motor vehicles, taxes on medicine, safety of medicine and potable water, discussion about tax police and abolition of tax police	B+
86	8	2	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)

130	1	1	CNN: fish kills, government water and environmental policy, Russian: world-wide problems with pure water supply, sickness, wars, fish kills	A
170	1	1	CNN: Mars exploration, water on Mars, radiation danger, Russian: Mars exploration, water on Mars, radiation danger, Mars Odyssey, “HEND” Mars mission	A+

Table A-2. English-Russian Clustering Performance - Russian Proper Names Amplified x 2

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
12	4	5 same as baseline	Academy Awards, Scorcese's Oscar Prospects, Others Oscar Prospects, Oscars during war in Iraq	A
27	3	1 lost 3 Russian Articles (MN)	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
47	1	14 same as baseline	CNN - senate/congress budget deliberations including tax cut; Russian: tax on automobiles and other motor vehicles, taxes on medicine, safety of medicine and potable water, discussion about tax police and abolition of tax police	B+
86	8	2 same as baseline	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)
104	2	3	CCN: Year in review, Chinese objections to was in Iraq: Russian: Far-east oil production, some mention of Iraq oil	F

108	2	11	CNN: New Microsoft products/technology, New PC products, Russian: Arts, theatre, critics, list of top women in industry (including 1 Microsoft exec.), Use of Micosoft products in Universities, interview with Microsoft woman exec.	C-
130	1	1 same as baseline	CNN: fish kills, government water and environmental policy, Russian: world-wide problems with pure water supply, sickness, wars, fish kills	A
170	1	1 same as baseline	CNN: Mars exploration, water on Mars, radiation danger, Russian: Mars exploration, water on Mars, radiation danger, Mars Odyssey, “HEND” Mars mission	A+
263	1	3	CNN: University President, wife, family arrested for drug offense, Russian: Refugees from Iraq, Iran provides humanitarian assistance	F

Table A-3. English-Russian Clustering Performance - Russian Proper Names Amplified x 3

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
8	2	1	CNN: Missing British troops, Iraqis surrender platoon, battles around UMM Qasar and Faw penisula, 16 british dead, Tornado pilots killed Russian: Abdication of Edward VIII, Canada's Role	F
10	1	1	CNN: Australian forces participation in war in Iraq, special forces, reconnaissance, Australian air combat, UM Qasar and humanitarian aid: Russian: Light-hearted treatment of US-French squabbling over war in Iraq, 'liberty fries' and French response	D-
27	3	1 lost 3 Russian Articles (MN)	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
86	8	2 same as baseline	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)
104	2	4	CCN: Year in review, Chinese objections to was in Iraq: Russian: Far-east oil production, some mention of Iraq oil	F
133	7	1	Computer viruses, SARS virus and its spread,	A

134	1	1	CNN: Colorado snow storms, insurance claims, effects on agriculture, fire suppression Russian: Hockey teams and players, especially Russian-American players	D+
161	1	1	CNN: Warnings against travel to Columbia, Russian: Bicycle races, sports	F
198	1	1	CNN: Health problems, obesity, Russian: Henry James plays	F
220	1	1	CNN: Gay rights, Russian: Anti-war comments by Dixie Chicks personage	F
241	1	1	CNN: Compendium of new military equipment, Russian: Helicopter crash in Chechnya	C-

Table A-4. English-Russian Clustering Performance - Russian Proper Names Amplified x 2 and CNN Proper Names Amplified x 2

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
12	6 two more than baseline	5 same as baseline	Academy Awards, Scorcese's Oscar Prospects, Related Coverage,Others Oscar Prospects, Oscars during war in Iraq	A
14	8	4	War in Iraq, Northern Iraq, Turkish reservations about Kurds, Stategic/Economic importance of Oil in North Iraq/Kirkut, Kurdish views on idependence, Ansar-Al-Islam-Al Queda link, reporter's diary account	A
27	3	1 lost 3 Russian Articles (MN)	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
84	1	1	CNN: Art, Michaelangelo, Russian: Abdication of Edward VIII, Canada's Role in abdication decision	F
86	8	1 lost 1 article	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)
104	2	3	CCN: Year in review, Chinese objections to was in Iraq: Russian: Far-east oil production, some mention of Iraq oil	F

108	2	11	CNN: New Microsoft products/technology, New PC products, Russian: Arts, theatre, critics, list of top women in industry (including 1 Microsoft exec.), Use of Micosoft products in Universities, interview with Microsoft woman exec.	C-
133	7	1	Computer viruses, SARS virus and its spread,	A *
161	1	1	CNN: Warnings against travel to Columbia, Russian: Bicycle races, sports	F
169	1	1	CNN: English mummies, Russian: Business with Egypt	D+
170	0	0 0	ABSENT	F
198	1	1	CNN: Health problems, obesity, Russian: Henry James plays	F
220	1	1	CNN: Gay rights, Russian: Anti-war comments by Dixie Chicks personage	F
241	1	1	CNN: Compendium of new military equipment, Russian: Helicopter crash in Chechnya	C-
262	1	1	CNN: Humanitarian assistance to Iraqi refugees, Russian: Brittain's refugee/immigration policies	C+

Table A-5. English-Russian Clustering Performance - Articles Translated by Babelfish

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
3	22	18 (-12)	CNN: War in Iraq, coalition casualties, Saddam TV appearance, POWs, Saddam Bio., weapons of mass destruction, Russian: War in Iraq, coalition and Iraqi causalties, POWs, Iraqi TV coverage, Iraqi response to bombing, friendly fire incidents/Patriot missiles, search for weapons of mass destruction, Iraqi TV propaganda, battle for Baghdad	A
6	2	3	CNN: UN aid to Iraq, Oil for Food Program, Blair - efforts to clear way for humanitarian aid, water supply, refugees fleeing to neighboring countries, Russian: British #1 for political/other refugees, humanitarian 'safe' countries defined by Britain, Iraqi refugees, humanitarian assistance to Iraqi refugees, Iranian border refugees	B-
12	5	5 (2 different)	Academy Awards, Scorcese's Oscar Prospects, Others Oscar Prospects, Oscars during war in Iraq	A
14	6	5 (+1)	War in Iraq, Northern Iraq, Turkish reservations about Kurds, Stategic/Economic importance of Oil in North Iraq/Kirkut, Kurdish views on idependence, Ansar-Al-Islam-Al Queda link	A

21	**4**	**3**	CNN: Nigerian Oil problems/related violence, Oil prices relating to Nigeria and Iraq, Venezuelan Oil prices, Chinese opposition to war in Iraq Russian: Russian Oil industry/market, Russian-Middle East oil connection, Proposed Russian-Chinese pileline, Russian-Chinese relations, Russian-Japanese relations and oil	A-
27	**3**	**3** (-1)	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
31	**3**	**1**	CNN: Impending battle for Baghdad, Republican guard strength/positions, US-British meetings, justification for war, British parliament, EU deliberations, personalities of Bush and Blair, US-EU-British relations Russian: Variety of stories including abolition of tax police, Russian tax policy, war in Iraq, Blair opinions on war, British opposition to war	B

47	**1**	**10** (-4)	CNN - senate/congress budget deliberations including tax cut; Russian: tax on automobiles and other motor vehicles, taxes on medicine, safety of medicine and potable water, discussion about tax police and abolition of tax police	B+
52	**3** (1 different)	**2** (-2)	CNN: Network and Cable TV losses due to Iraq war coverage, NY Times/Discovery Station merged channel including mention of stories related to terrorism and to Iraq Russian: Links to Moscow 'Year in Review' stories	C-
67	**1**	**6** (+1)	CNN: Troublesome dogs helped by Japanese 'Bowlingual' product that translates soothing messages to dogs, in dog language. Russian: Attacks by vicious dogs, Novosibersk and other Russian locales, Russian/Regional laws pertaining to dog attacks and leash laws, Case studies in vicious dog attacks	A-

75	**1**	**1**	CNN: Monical Lewinsky selected by Fox network to emcee ‘Mr. Personality’ - a hidden-identity dataing game. Russian: Alleged Bill Clinton affair with Russian supermodel, Clinton psychological sketch relating to sexual misadventures, Hillary’s presidential prospects	A
86	**7** (-1)	**2**	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)
130	**1**	**2** (+1)	CNN: fish kills, government water and environmental policy, Russian: world-wide problems with pure water supply, sickness, wars, fish kills	A
138	**5**	**1**	CNN: Two helicopter pilots taken prisoner in Iraq, US Military hardware in Iraq including several helicopters Russian: Two helicopters crash in Chechnya, three versions of accident	A-
170	**1**	**1**	CNN: Mars exploration, water on Mars, radiation danger, Russian: Mars exploration, water on Mars, radiation danger, Mars Odyssey, “HEND” Mars mission	A+

188	1 (prev. part of 86)	1 (prev. part of 396)	CNN: US urged to take lead in global disease prevention, pneumonia, SARS, influenza, prevention, new pathogens, antibiotics, antivirulents Russian: Enzyme therapy, influenza, immunomodulation, various immuotherary products	B
202	**1**	**3** (-2)	CNN: Officals warn against poisoning danger for kids, child-resistant packaging, household drugs, proper dosage Russian: Russian medical and prescription benefits, taxes on prescriptions and healthcare, health tax legislation, higher Russian mortality rates and reasons	C
218	**1**	**1** (-2)	CNN: Parents on trial for child's severe malnutrition: Russian: Russian polar expeditions, including those kept secret during the Cold War, Northern Arctic, medical experiments conducted	D

232	3 (-1)	5 (-2)	CNN: US accuses Russia of supplying forbidden arms to Iraq including GPS naviational jammers and night vision equipment. Russian: U.S. State Dept. accuses Russia of providing military supplies to Iraq including GPS navigational jammers and night-vision equiptment. Russian denies allegations, Russian economic relations with Iraq including selling armaments and anti-tank missiles and equipment, US Star Wars, Strategic Defense Initiative and Russian diplomatic and political response	A+
259	1 (prev. part of 258)	3	CNN: Schools aim to reassure students during wartime. Russian: Bribery and corruption in Russian educational system, students abused by teachers, homeless children and attempts by government and individuals to help them	B+

note: Clusters in RED appear both in runs using Babelfish translation and dictionary-based translations

Numbers in GREEN indicate monolingual clusters are the same for both runs

Numbers in parentheses indicate number of discrepancies in monolingual clusters, otherwise the same for both runs.

Table A-6. English-Russian Clustering Performance - Word-for-Word After Training

Cluster #	# CNN Articles	# Russian Articles	Topics	Quality
7	**10**	**4**	CNN: Bush's motivation for war in Iraq, world and UN response to US actions, search for weapons of mass destruction, Iraq connection to terrorism, neo-conservative views on US world role, Bush gives estimate of war's cost, voices concern over Russia's lack of support, America pressures Russia over its continuous military aid to Iraq. Russian: Iraq-Bin Laden connection, War in Iraq, G. Bush address justifying war based on suspected Iraqi Weapons of Mass Destruction, Chem/Bio weapons, Iraq missiles' ability to strike Saudi Arabia and neighbors, Russia/France stalling tactics, elder Bush's sentiments about 1991 war in Iraq and comments about present conflict.	A-
12	**5** (-1)	**6** (+1)	Academy Awards, Scorcese's Oscar Prospects, Others Oscar Prospects, Oscars during war in Iraq	A
14	**6**	**4**	War in Iraq, Northern Iraq, Turkish reservations about Kurds, Stategic/Economic importance of Oil in North Iraq/Kirkut, Kurdish views on idependence, Ansar-Al-Islam-Al Queda link	A

21	3	1	CNN: Nigerian Oil problems/related violence, Oil prices relating to Nigeria and Iraq, Venezuelan Oil prices, Chinese opposition to war in Iraq Russian: Russian Oil industry/market, Russian-Middle East oil connection, Proposed Russian-Chinese pileline, Russian-Chinese relations, Russian-Japanese relations and oil	A-
27	3	4	CNN - List of countries vs. currencies, Russian - rumors about Iraqi Dinar, American Dollar also discussion of monetary exchange rates and trends including dollar, ruble, euro and others	A-
31	3	1	CNN: Impending battle for Baghdad, Republican guard strength/positions, US-British meetings, justification for war, British parliament, EU deliberations, personalities of Bush and Blair, US-EU-British relations Russian: Variety of stories including abolition of tax police, Russian tax policy, war in Iraq, Blair opinions on war, British opposition to war	B

47	**1**	**14**	CNN - senate/congress budget deliberations including tax cut; Russian: tax on automobiles and other motor vehicles, taxes on medicine, safety of medicine and potable water, discussion about tax police and abolition of tax police	B+
86	**8**	**2**	Computer viruses, SARS virus and its spread,	A * (computer and human viruses treated the same)
104	2	4	CCN: Year in review, Chinese objections to was in Iraq: Russian: Far-east oil production, some mention of Iraq oil	F
130	**1**	**1**	CNN: fish kills, government water and environmental policy, Russian: world-wide problems with pure water supply, sickness, wars, fish kills	A
156	**1**	**1**	CNN: Hawaiian farmers welcome tourists, agricultural tourism, legislation regulating Hawaiian agricultural tourism, tourist destinations. Russian: Moscow International Travel Expo, travel to foreign countries including Turkey, Egypt, Cyprus, Spain as well as more exotic locales. Outlook for Russian travel industry.	A-

170	**1**	**1**	CNN: Mars exploration, water on Mars, radiation danger, Russian: Mars exploration, water on Mars, radiation danger, Mars Odyssey, “HEND” Mars mission	A+

note: Clusters in BLUE appear both in runs using the trained dictionary-based translation and Babelfish translation

Numbers in GREEN indicate monolingual clusters are the same for both runs

Numbers in parentheses indicate number of discrepancies in monolingual clusters, otherwise the same for both runs