Two Statistical Summarizers at INEX 2012 Tweet Contextualization Track

Juan-Manuel Torres-Moreno[1,2] and Patricia Velázquez-Morales[3]

[1] École Polytechnique de Montréal - Département de Génie Informatique
CP 6079 Succ. Centre Ville H3C 3A7 Montréal (Québec), Canada.
[2] Université d'Avignon et des Pays de Vaucluse
BP 911228, 84911 Avignon Cedex 9, France
juan-manuel.torres@univ-avignon.fr
[3] VM Labs, 84000 Avignon, France.
patricia_velazquez@yahoo.com
http://www.polymtl.ca

Abstract. According to the organizers, the objective of the 2012 INEX Tweet Contextualization Task is: "*...given a tweet, the system must provide some context about the subject of the tweet, in order to help the reader to understand it. This context should take the form of a readable (and short) summary, composed of passages from [...] Wikipedia.*" We present summarizers Cortex and KL-summ applied to the INEX 2012 task. Cortex summarizer uses several sentence selection metrics and an optimal decision module to score sentences from a document source. KL-summ is a new statistical summarizer based on Kullback-Leibler divergence (the same used by INEX organizers) to score sentences. The results show that Cortex system (using original tweets) outperforms KL-summ on INEX task.

Keywords: INEX, Automatic Summarization System, Tweet contextualization, Cortex, KL Divergence.

1 Introduction

Automatic text summarization is indispensable to cope with ever increasing volumes of valuable information. An abstract is by far the most concrete and most recognized kind of text condensation [1, 2]. We adopted a simpler method, usually called *extraction*, that allow to generate summaries by extraction of pertinence sentences [2–5]. Essentially, extracting aims at producing a shorter version of the text by selecting the most relevant sentences of the original text, which we juxtapose without any modification. The vector space model [6, 7] has been used in information extraction, information retrieval, question-answering, and it may also be used in text summarization [8]. CORTEX[4] is an automatic

[4] *CORTEX es Otro Resumidor de TEXtos* (CORTEX is anotheR TEXt summarizer).

summarization system [9] which combines several statistical methods with an optimal decision algorithm, to choose the most relevant sentences.

An open domain Question-Answering system (QA) has to precisely answer a question expressed in natural language. QA systems are confronted with a fine and difficult task because they are expected to supply specific information and not whole documents. At present there exists a strong demand for this kind of text processing systems on the Internet. A QA system comprises, *a priori*, the following stages [10]:

- Transform the questions into queries, then associate them to a set of documents;
- Filter and sort these documents to calculate various degrees of similarity;
- Identify the sentences which might contain the answers, then extract text fragments from them that constitute the answers. In this phase an analysis using Named Entities (NE) is essential to find the expected answers.

Most research efforts in summarization emphasize generic summarization [11–13]. User query terms are commonly used in information retrieval tasks. However, there are few papers in literature that propose to employ this approach in summarization systems [14–16]. In the systems described in [14], a learning approach is used (performed). A document set is used to train a classifier that estimates the probability that a given sentence is included in the extract. In [15], several features (document title, location of a sentence in the document, cluster of significant words and occurrence of terms present in the query) are applied to score the sentences. In [16] learning and feature approaches are combined in a two-step system: a training system and a generator system. Score features include short length sentence, sentence position in the document, sentence position in the paragraph, and tf.idf metrics. Our generic summarization system includes a set of eleven independent metrics combined by a Decision Algorithm. Query-based summaries can be generated by our system using a modification of the scoring method. In both cases, no training phase is necessary in our system.

This paper is organized as follows. In Section 2 we explain the INEX 2012 Tweet Contextualization Track. In Section 3 we explain the methodology of our work. Experimental settings and results obtained with Cortex and KL-summ summarizers are presented in Section 4. Section 5 exposes the conclusions of the paper and the future work.

2 INEX 2012 Tweet Contextualization Track

The Initiative for the Evaluation of XML Retrieval (INEX) is an established evaluation forum for XML information retrieval (IR) [17]. In 2012, tweet contextualization INEX task at CLEF 2012, aims "*given a new tweet, the system must provide some context about the subject of the tweet, in order to help the reader to understand it. This context should take the form of* **a readable summary, not exceeding 500 words, composed of passages from a provided Wikipedia corpus.**"[5]

[5] https://inex.mmci.uni-saarland.de/tracks/qa/

Like in Question Answering track of INEX 2011, the present task is about contextualizing tweets, i.e. answering questions of the form "What is this tweet about?" using a recent cleaned dump of the Wikipedia[6]. As organizers claim, the general process involves three steps:

- Tweet analysis.
- Passage and/or XML elements retrieval.
- Construction of the answer.

Then, a relevant passage segment contains:

- Relevant information but
- As few non-relevant information as possible (the result is specific to the question).

2.1 Document Collection

The corpus has been constructed from a dump of the English Wikipedia from November 2011. All notes and bibliographic references were removed to facilite the extraction of plain text answers. (Notes and bibliographic references are difficult to handle). Resulting documents contains a title, an abstract and section. Each section has a sub-title. Abstract end sections are made of paragraphs and each paragraph can have entities that refer to Wikipedia pages.

2.2 Tweets set

The committee of INEX has defined about 1000 tweets for the Track 2012. 1133 tweets in English were collected by the organizers from Twitter[7]. Tweets were selected and checked among informative accounts (for example, @CNN, @TennisTweets, @PeopleMag, @science...), in order to avoid purely personal tweets that could not be contextualized. Information such as the user name, tags or URLs will be provided.

3 The Text Summarizers used

3.1 Cortex Summarization System

Cortex [18, 19] is a single-document extract summarization system. It uses an optimal decision algorithm that combines several metrics. These metrics result from processing statistical and informational algorithms on the document vector space representation.

The INEX 2012 Tweet Contextualization Track evaluation is a real-world complex question (called long query) answering, in which the answer is a summary constructed from a set of relevant documents. The documents are parsed

[6] See the official INEX 2012 Tweet Contextualization Track Website: `https://inex.mmci.uni-saarland.de/tracks/qa/`.

[7] `www.tweeter.com`

to create a corpus composed of the query and the the multi-document retrieved by a Perl program supplied by INEX organizers[8]. This program is coupled to Indri system[9] to obtain for each query, 50 documents from the whole corpus.

The idea is to represent the text in an appropriate vectorial space and apply numeric treatments to it. In order to reduce complexity, a preprocessing is performed to the question and the document: words are filtered, lemmatized and stemmed.

The Cortex system uses 11 metrics (see [20, 19] for a detailed description of these metrics) to evaluate the sentence's relevance.

1. The frequency of words.
2. The overlap between the words of the query (R).
3. The entropy the words (E).
4. The shape of text (Z).
5. The angle between question and document vectors (A).
6. The sum of Hamming weigths of words per segment times the number of different words in a sentence.
7. The sum of Hamming weights of the words multiplied by word frequencies.
8. The words interaction (I).
9. ...

By example, the topic-sentence overlap measure assigns a higher ranking for the sentences containing question words and makes selected sentences more relevant. The overlap is defined as the normalized cardinality of the intersection between the query word set T and the sentence word set S.

$$\text{Overlap}(T, S) = \frac{card(S \cap T)}{card(T)} \tag{1}$$

The system scores each sentence with a decision algorithm that relies on the normalized metrics. Before combining the votes of the metrics, these are partitioned into two sets: one set contains every metric $\lambda^i > 0.5$, while the other set contains every metric $\lambda^i < 0.5$ (values equal to 0.5 are ignored). We then calculate two values α and β, which give the sum of distances (positive for α and negative for β) to the threshold 0.5 (the number of metrics is Γ, which is 11 in our experiment):

$$\alpha = \sum_{i=1}^{\Gamma} (\lambda^i - 0.5); \quad \lambda^i > 0.5 \tag{2}$$

$$\beta = \sum_{i=1}^{\Gamma} (0.5 - \lambda^i); \quad \lambda^i < 0.5 \tag{3}$$

The value given to each sentence s given a query q is calculated with:

[8] See: `http://qa.termwatch.es/data/getINEX2011corpus.pl.gz`

[9] Indri is a search engine from the Lemur project, a cooperative work between the University of Massachusetts and Carnegie Mellon University in order to build language modelling information retrieval tools. See: `http://www.lemurproject.org/indri/`

$$\begin{aligned} &\text{if}(\alpha > \beta) \\ &\qquad \text{then } \ \text{Score}(s,q) = 0.5 + \tfrac{\alpha}{\Gamma} \\ &\qquad \text{else } \ \text{Score}(s,q) = 0.5 - \tfrac{\beta}{\Gamma} \end{aligned} \tag{4}$$

The Cortex system is applied to each document of a topic and the summary is generated by concatenating higher score sentences.

3.2 The KL-summ summarization system

The main idea of KL-summarizer is to weight the sentences of a document, by minimizing the divergence of each sentence from document source. This idea is quite simple. Several divergence measures can be utilized: Jensen-Shannon, Kullback-Leibler, etc. However, in order to obtain a good summarizer on this specific task, we decide of implement the same measure of evaluation proposed by the INEX' organizers.

Fresa measure [21, 22] is similar to Rouge evaluation [23] but it does not uses reference summaries. It calculates the divergence of probabilities between the candidate summary and the document source. Among these metrics, Kullback-Leibler (KL) and Jensen-Shannon (JS) divergences have been used [24, 21] to evaluate the informativeness of summaries. In this paper, we use Fresa, based in KL divergence with Dirichlet smoothing, like in the 2010, 2011 and 2012 INEX edition [25], to evaluate the informative content of summaries by comparing their n-gram distributions with those from source documents.

Fresa simply considered absolute log-diff between frequencies. Let T be the set of terms in the source. For every $t \in T$, we denote by C_t^T its occurrences in the source and by C_t^S its occurrences in the summary. The Fresa package computed the divergence between source and summaries as:

$$\mathcal{D}(T||S) = \sum_{t \in T} \left| \log\left(\frac{C_t^T}{|T|} + 1\right) - \log\left(\frac{C_t^S}{|S|} + 1\right) \right| \tag{5}$$

To score each sentence, several automatic measures were computed:

- Fresa_1: Uni-grams of single stems after removing stop-words.
- Fresa_2: Bi-grams of pairs of consecutive stems (in the same sentence).
- Fresa_{SU4}: Bi-grams with 2-gaps also made of pairs of consecutive stems but allowing the insertion between them of a maximum of two stems.

All Fresa scores, $\text{Fresa}_1 = 1 - \mathcal{D}(T||S)$ are normalized between 0 and 1. High values mean a less divergence of summary from source document. In other words, lower divergences (High Fresa scores) shows a more quantity of content of summary.

So, the relevant sentences will be selected as having the less divergence values. The two first modules are based on the Cortex system[10]. Finally, the third

[10] See section 3.1.

module generates summaries by displaying and concatenating of the relevant sentences.

At first, the first 50 documents of the cluster are concatenated into a single multi-document in chronological order. Placing the tweet q (enriched or not) like the title of this long document. The divergence between each sentence among the all others is computed using equation 5.

4 Experiments Settings and Results

In this study, we used the document sets made available during the Initiative for the Evaluation of XML retrieval (INEX)[11], in particular on the INEX 2012 Tweet Contextualization Track.

The strategy of Cortex and KL-summ systems to deal multi-document summary problem is quite simple: first, a long single document D is formed by concatenation of all $i = 1, ..., n$ relevant documents provided by Indri engine: $d_1, d_2, ...d_n$. The first line of this multi-document D is the tweet T. Both summarizers systems extract of D the most relevant sentences following T. Then, this subset of sentences is sorted by the date of documents d_i. The summarizers add sentences into the summary until the word limit is reached. To evaluate the performance of Cortex and KL-summ systems on INEX tweet contextualization track, we used the online package available from INEX website[12].

4.1 INEX Tweets enrichment

Two different strategies were employed to generate 1133 queries from tweets:

1. No pre-processing of tweet.
2. Enrichment of each tweet by semi-automatic synonyms of 100 heavy terms (their weights were calculated using the tf).

1) No pre-processing or modification was applied on queries set. Summarizers use the query as a title of a big multi-document retrieved by Indri engine.

2) Enrichment of tweet. The query has been semi-manually enriched as following. Firstly, a list T terms from the set of tweets was extracted then sorted by their term frequency. Liste T was manually inspected to extract the 100 first "relevant" terms. These 100 relevant terms were injected into tweets to enriched them.

Table 1 shows an example of the results obtained by Cortex and KL-summ systems using the 50 first documents retreived by Indri as input. The tweet that the summary should contextualiser in this case was the number 169231181181747200:

```
<topic id="169231181181747200">
<tweet>
```

[11] https://inex.mmci.uni-saarland.de/

[12] http://qa.termwatch.es/data/

```
CNNLive View stake-out camera funeral home WhitneyHouston body
expected arrive New Jersey Watch live
</tweet>
```

The tweets were enriched automatically using the 100 first terms (sorted by their tf weight) obtained from the words contained in the tweets set.

For example, for the tweet 169231181181747200, strategy 2 produce the following list of synonyms:

```
-- funeral : cremation
-- home : domicile
-- new : recent
-- watch : observe
-- live : exist
```

Then, query 169231181181747200 is enriched as show:
q = "`cnnlive view stakeout camera funeral cremation home domicile whitneyhouston body expected arrive new recent jersey watch observe live exist`"

Compound words are not detected in this phase. Since, "New Jersey" was separated in "New" and "Jersey", then "New" was enriched by "recent". This criterion is simple but it seems well enrich some tweets too shorts.

Table 1 presents Cortex and KL-summ results (queries enriched or not; tweet=167355997915058176 see Appendix) in comparison with the INEX baseline (Baseline summary), and three baselines, that is, summaries including random *n*-grams (Random uni-grams) and 5-grams (Random 5-grams) and empty baseline. In this particular example, we observe that KL-summ outperforms all summarizers.

Table 1. Example of Summarization results on tweet 167355997915058176.

Summary type	Uni-grams	Bi-grams	SU4 bi-grams	FRESA Averages
Baseline summary	33.36151	41.42226	41.40674	38.73017
Empty baseline	45.13673	53.58049	53.52617	50.74779
Random uni-grams	35.10965	43.58796	43.50809	40.73523
Random 5-grams	31.52959	39.73750	39.76517	37.01075
Cortex (query=Tweet)	32.16833	40.07052	40.14011	*37.45966*
Cortex (query=Tweet+Synonyms)	33.33823	41.26795	41.33727	38.64782
KL-summ (Query=Tweet)	31.79170	39.66283	39.74201	**37.06551**

Figure 1 shows the official results of participants of INEX 2012 contextualization task. The performances (rank) of our summarizers are: Cortex=11/33, Cortex with enriched tweets=21/33 and KL-summ=16/33. In this figure, rank axes represents the SU4-bigrams values.

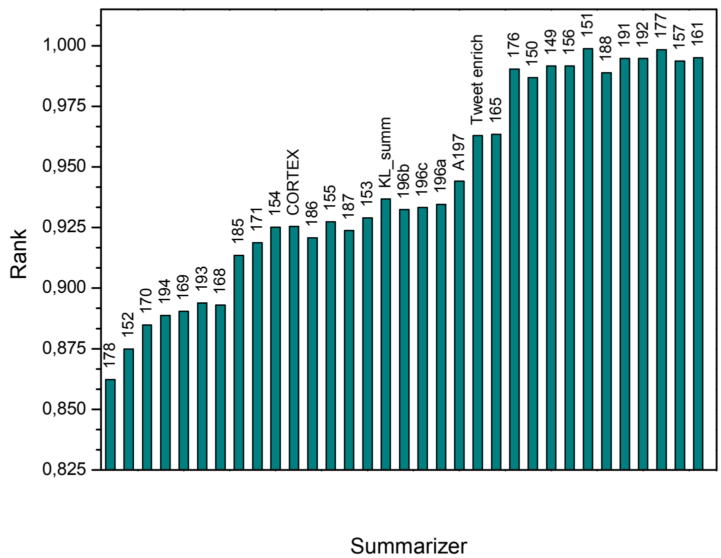

Fig. 1. Official results for systems participants on INEX 2012 contextualization task. Our systems are: CORTEX, KL-summ and CORTEX with enriched tweets.

5 Conclusions

In this paper we have presented the Cortex and KL-summ summarization systems applied on INEX 2012 Tweet Contextualization Track. The first one is based on the fusion process of several different sentence selection metrics. The decision algorithm obtains good scores on the INEX 2012 Tweet Contextualization Track (the decision process is a good strategy without training corpus). The second one is based on the divergence between summary and the source document.

Cortex summarizer using original tweets as inputs has obtained very good results in the automatic FRESA evaluations. In fact, semi-automatic tweet enrichment has disappointed in this task. Cortex using original tweets outperforms Cortex using enriched queries. We think that the strategy of enrichment, without compound words detection, was a very simple process. A module of compound words may improves the performance of this strategy. In other hands, the KL-summ summarizer is slightly less good than Cortex system. In fact, function 5 is not additive. Since, a simple sort based on Fresa scores (1-divergence calculated in local form) is not enough to maximize the global score over the whole document. A more complex strategy of optimization must be implemented in order to deal this problem. We show that simple statistical summarizers show good performances in this complex task.

References

1. ANSI. American National Standard for Writing Abstracts. Technical report, American National Standards Institute, Inc., New York, NY, 1979. (ANSI Z39.14.1979).
2. J.M. Torres-Moreno. *Resume automatique de documents : une approche statistique*. Hermes-Lavoisier, 2011.
3. H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159, 1958.
4. H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.
5. I. Mani and M. Mayburi. *Advances in automatic text summarization*. The MIT Press, U.S.A., 1999.
6. Gregory Salton. *The SMART Retrieval System - Experiments un Automatic Document Processing*. Englewood Cliffs, 1971.
7. Gregory Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
8. I. Da Cunha, S. Fernandez, P. Velázquez Morales, J. Vivaldi, E. SanJuan, and J.M. Torres-Moreno. A new hybrid summarizer based on vector space model, statistical physics and linguistics. In *MICAI 2007: Advances in Artificial Intelligence*, pages 872–882. Springer Berlin/Heidelberg, 2007.
9. J.M. Torres-Moreno, P. Velazquez-Morales, and JG. Meunier. Condensés automatiques de textes. *Lexicometrica. L'analyse de données textuelles : De l'enquête aux corpus littéraires*, Special(www.cavi.univ-paris3.fr/lexicometrica), 2004.
10. C. Jacquemin and P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. *Le document en sciences du traitement de l'information*, 4:71–109, 2000.
11. Jose Abracos and Gabriel Pereira Lopes. Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles. In Inderjeet Mani and Mark T. Maybury, editors, *ACL/EACL97-WS*, Madrid, Spain, July 11 1997.
12. Simone Teufel and Marc Moens. Sentence Extraction as a Classification Task. In Inderjeet Mani and Mark T. Maybury, editors, *ACL/EACL97-WS*, Madrid, Spain, 1997.
13. Eduard Hovy and Chin Yew Lin. Automated Text Summarization in SUMMARIST. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. The MIT Press, 1999.
14. Julian Kupiec, Jan O. Pedersen, and Francine Chen. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, 1995.
15. Anastasios Tombros, Mark Sanderson, and Phil Gray. Advantages of Query Biased Summaries in Information Retrieval. In Eduard Hovy and Dragomir R. Radev, editors, *AAAI98-S*, pages 34–43, Stanford, California, USA, March 23–25 1998. The AAAI Press.

16. Judith D. Schlesinger, Deborah J. Backer, and Robert L. Donway. Using Document Features and Statistical Modeling to Improve Query-Based Summarization. In *DUC'01*, New Orleans, LA, 2001.
17. Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman, editors. *Comparative Evaluation of Focused Retrieval - 9th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2010, Vugh, The Netherlands, December 13-15, 2010, Revised Selected Papers*, volume 6932 of *Lecture Notes in Computer Science*. Springer, 2011.
18. J.M. Torres-Moreno, P. Velazquez-Moralez, and J. Meunier. *CORTEX, un algorithme pour la condensation automatique de textes*. In *ARCo*, volume 2, page 365, 2005.
19. J.M. Torres-Moreno, P.L. St-Onge, M. Gagnon, M. El-Bèze, and P. Bellot. Automatic summarization system coupled with a question-answering system (qaas). *in CoRR*, abs/0905.2990, 2009.
20. J.M. Torres-Moreno, P. Velazquez-Morales, and J.G. Meunier. *Condensés de textes par des méthodes numériques*. *JADT*, 2:723–734, 2002.
21. J.-M. Torres-Moreno, H. Saggion, I. da Cunha, E. SanJuan, and P. Velázquez-Morales. Summary Evaluation with and without References. *Polibits*, 42(42):13–19, 2010.
22. H. Saggion, J.-M. Torres-Moreno, I. da Cunha, and E. SanJuan. Multilingual summarization evaluation without human models. In *23rd Int. Conf. on Computational Linguistics*, COLING '10, pages 1059–1067, Beijing, China, 2010. ACL.
23. C.-Y. Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out*, 2004.
24. Annie Louis and Ani Nenkova. Automatically Evaluating Content Selection in Summarization without Human Models. In *Empirical Methods in Natural Language Processing*, pages 306–314, Singapore, August 2009.
25. Eric SanJuan, Patrice Bellot, Véronique Moriceau, and Xavier Tannier. Overview of the inex 2010 question answering track (qa@inex). In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman, editors, *Comparative Evaluation of Focused Retrieval*, volume 6932 of *Lecture Notes in Computer Science*, pages 269–281. Springer Berlin / Heidelberg, 2011.

6 Appendix

We present the Cortex summary of tweet 167355997915058176: *"Scientists claim 42000-year-old paintings of seals by Neanderthals found in Spanish cave CIZy-Rotb"*. (Numbers in bold indicate the weight of each sentence).

0.861 Therefore, as human populations slowly increased, the cave bear faced a shrinking pool of suitable caves, and slowly faded away to extinction, as both Neanderthals and anatomically modern humans sought out caves as living quarters, depriving the cave bear of vital habitat. **0.880** The evidence found in archeological sites suggests these early humans were Nomadic, that they lived in caves, and acquired sustenance by hunting wild boar, red deer, mountain goat s, fallow deer, and horse s, competing with other predators like the leopard, the brown bear, and the wolf. **0.902** Late Oldowan/Early Acheulean humans such as "Homo ergaster/Homo erectus" may have been the first people to invent central campsites or home bases and incorporate them into their foraging and hunting strategies like contemporary hunter-gatherers, possibly as early as 1.7 million years ago; however, the earliest solid evidence for the existence of home bases or central campsites among humans only dates back to 500,000 years ago. **0.859** Christopher Boehm Harvard university press Raymond C. Kelly speculates that the relative peacefulness of Middle and Upper Paleolithic societies resulted from a low population density, cooperative relationships between groups such as reciprocal exchange of commodities and collaboration on hunting expeditions, and because the invention of projectile weapons such as throwing spears provided less incentive for war, because they increased the damage done to the attacker and decreased the relative amount of territory attackers could gain. **1.000** In particular, Emil Bächler suggested that a bear cult was widespread among Middle Paleolithic Neanderthal s. A claim that evidence was found for Middle Paleolithic animal worship c 70,000 BCE originates from the Tsodilo Hills in the African Kalahari desert has been denied by the original investigators of the site. **0.865** However, recent archaeological research done by the anthropologist and archaeologist Steven Kuhn from the University of Arizona reveals that this gender-based division of labor did not exist prior to the Upper Paleolithic in Middle Paleolithic societies and was invented relatively recently in human prehistory. **0.871** One theory holds that behavioral modernity occurred as a sudden event some 50 kya in prehistory, possibly as a result of a major genetic mutation or as a result of a biological reorganization of the brain that led to the emergence of modern human natural language s. Proponents of this theory refer to this event as the "Great Leap Forward" "or the" "Upper Paleolithic Revolution". **0.941** Since genetics does not reject the hypothesis of a Neanderthal-modern admixture, and morphological and archaeological evidence suggest that Neanderthal lineages survived into later Upper Paleolithic populations, Pestera cu Oase findings provide a strong argument in favor of an admixture model between regional Neanderthals and early modern humans. **0.965** In another study, researchers have recently found in Pestera Muierilor, Romania, remains of European humans from years ago who possessed mostly diagnostic modern anatomical features, but "also" had distinct Neanderthal features not present in ancestral modern humans in Africa, including a large bulge at the back of the skull, a more prominent projection around the elbow joint, and a narrow socket at the shoulder joint. **0.910** The distribution of the D allele, high outside Africa but low in sub-Saharan Africa, has been suggested to indicate involvement of an archaic Eurasian population, and current estimates of the divergence time between modern humans and Neanderthals based on mitochondrial DNA are in favor of the Neanderthal lineage as the most likely archaic Homo population from which introgression into the modern human gene pool took place.