Estimating Language Relationships from a Parallel Corpus. A Study of the Europarl Corpus

Taraka Rama, Lars Borin
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
taraka.rama.kasicheyanula@gu.se
lars.borin@svenska.gu.se

Abstract

Since the 1950s, linguists have been using short lists (40–200 items) of basic vocabulary as the central component in a methodology which is claimed to make it possible to automatically calculate genetic relationships among languages. In the last few years these methods have experienced something of a revival, in that more languages are involved, different distance measures are systematically compared and evaluated, and methods from computational biology are used for calculating language family trees. In this paper, we explore how this methodology can be extended in another direction, by using larger word lists automatically extracted from a parallel corpus using word alignment software. We present preliminary results from using the Europarl parallel corpus in this way for estimating the distances between some languages in the Indo-European language family.

1 Introduction

Automatic identification of genetic relationships among languages has gained attention in the last few years. Estimating the distance matrix between the languages under comparison is the first step in this direction. Then a distance based clustering algorithm can be used to construct the phylogenetic tree for a family. The distance matrix can be computed in many ways. Lexical, syntactic and semantic features of the languages can be used for computing this matrix (Ringe et al., 2002). Of these, lexical features are the most widely used features, most commonly in the form of *Swadesh lists*.

Swadesh lists are short lists (40–200 items) of basic senses which are supposed to be universal. Further, the words expressing these senses in a language are supposed to be resistant to borrowing. If these two assumptions hold, it follows that such lists can be used to calculate a numerical estimate of genetic distances among related languages, an endeavor referred to as *lexicostatistics*. A third assumption which was often made in the older literature was that the replacement rate of this basic vocabulary was constant and could be expressed as a constant percentage of the basic vocabulary being replaced over some unit of time (exponential decay). This third assumption has generally been abandoned as flawed and with it the body of research that it motivated, often referred to as *glottochronology*.

In lexicostatistics, the similarity between two languages is the percentage of shared cognates between the two languages in such a list. In the terminology of historical linguistics, cognates are words across languages which have descended independently in each language from the same word in a common ancestor language. Hence, loanwords are not cognates. Cognates are identified through regular sound correspondences. For example, English ∼ German *night* ∼ *Nacht* ‘night’ and *hound* ∼ *Hund* ‘dog’ are cognates. If the languages are far enough removed in time, so that sound changes have been extensive, it is often far from obvious to the non-expert which words are cognates, e.g. English ∼ Greek *hound* ∼ *kuon* ‘dog’ or English ∼ Armenian *two* ∼ *erku* ‘two’.

In older lexicostatistical work (e.g. Dyen et al. 1992), cognates are manually identified as such by experts, but in recent years there has been a strong interest in developing automatic methods for cognate identification. The methods proposed so far are generally based on some form of orthographic similarity[1] and cannot distinguish be-

[1] Even though the similarity measures used in the literature all work with written representations of words, these written representations are often in fact phonetic transcriptions, so that we can say that we have a phonetic similarity measure. For this reason we will use “orthographic” and “phonetic” interchangeably below.

tween cognates on the one hand and loanwords or chance resemblances on the other. Confusingly, the word pairings or groups identified in this way are often called cognates in the computational linguistics literature, whereas the term *correlates* has been proposed in historical linguistics for the same thing (McMahon and McMahon, 2005). In any case, the identification of such orthographically similar words is a central component in any automatic procedure purporting to identify cognates in the narrower sense of historical linguistics. Hence, below we will generally refer to these methods as methods for the identification of cognates, even if they actually in most cases identify correlates.

There have been numerous studies employing string similarity measures for the identification of cognates. The most commonly used measure is normalized edit distance. It is defined as the minimum number of deletions, substitutions and insertions required to transform one string to another. There have also been studies on employing identification of cognates using string similarity measures for the tasks of sentence alignment (Simard et al., 1993), statistical machine translation (Kondrak et al., 2003) and translational lexicon extraction (Koehn and Knight, 2002).

The rest of this paper is structured as follows. Section 2 discusses related work. Section 3 explains the motivation for using a parallel corpus and describes the approach.

2 Related work

Kondrak (2002) compares a number of algorithms based on phonetic and orthographical similarity for judging the cognateness of a word pair. His work surveys string similarity/ distance measures such as *edit distance*, *Dice coefficient* and *longest common subsequence ratio* (LCSR) for the task of cognate identification. The measures were tested on vocabulary lists for the Algonquian language family and Dyen's (1992) Indo-European lists.

Many studies based on lexicostatistics and phylogenetic software have been conducted using Swadesh lists for different language families. Among the notable studies for Indo-European are the lexicostatistical experiments of Dyen et al. (1992) and the phylogeny experiments of Ringe et al. (2002) and Gray and Atkinson (2003). In another study, Ellison and Kirby (2006) used intra-language lexical divergence for measuring the inter-language distances for the Indo-European language family.

Recently, a group of scholars (Wichmann et al., 2010; Holman et al., 2008) have collected 40-item Swadesh word lists for about two thirds of the world's languages.[2] This group uses a modified Levenshtein distance between the lexical items as the measure of the inter-language distance.

Singh and Surana (2007) use corpus based measures for estimating the distances between South Asian languages from noisy corpora of nine languages. They use a phonetics based similarity measure called *computational phonetic model of scripts* (CPMS; Singh et al. 2007) for pruning the possible cognate pairs between languages. The mean of the similarity between the pruned cognate pairs using this measure is estimated as the distance between the languages.

Bergsma and Kondrak (2007) conduct experiments for cognate identification using alignment-based discriminative string similarity. They automatically extract cognate candidate pairs from the Europarl corpus (Koehn, 2005) and from bilingual dictionaries for the language pairs English–French, English–German, English–Greek, English–Japanese, English–Russian, and English–Spanish. Bouchard-Côté et al. (2007) also use the Europarl corpus to extract cognates for the task of modeling the diachronic phonology of the Romance languages. In neither case is the goal of the authors to group the languages genetically by family, as in the work presented here. The previous work which comes closest to the work presented here is that of Koehn (2005), who trains pair-wise statistical translation systems for the 11 languages of the Europarl corpus and uses the systems' BLEU scores for clustering the languages, under the assumption that ease of translation correlates with genetic closeness.

3 Our approach

As noted above, automatic identification of cognates is a crucial step in computational historical linguistics. This requires an approach in which cognates have to be identified with high precision. This issue has been discussed by Brew et al. (1996). They were trying to extract possi-

[2] Their collaboration goes under the name of the *Automated Similarity Judgement Program (ASJP)* and their current dataset (in late 2010) contains word lists for 4,820 languages, where all items are rendered in a coarse phonetic transcription, even for those languages where a conventional written form exists.

ble English-French translation pairs from a multilingual corpus for the task of computational lexicography. Two issues with the automatic methods is the presence of *false friends* and *false negatives*. False friends are word pairs which are similar to each other but are unrelated. Some examples of false friends in French and English are *luxure* 'lust' ∼ *luxury*; *blesser* 'to injure' ∼ *bless*. False negatives are word pairs which are actually cognates but were identified as unrelated. For our task, we focus on identifying cognates with a high precision – i.e., few false friends – and a low recall – i.e., many false negatives. The method requires that the word pairs are translations of each other and also have a high orthographic similarity.

Section 4 introduces the use of the Europarl corpus for cognate identification. We extract the cognate pairs between a pair of languages in the following manner. For every language pair, the corpus is word aligned using GIZA++ (Och and Ney, 2003) and the word pairs are extracted from the alignments. Word pairs with punctuation are removed from the final set. Positive and negative training examples are generated by thresholding with a LCSR cutoff of 0.58.

The cutoff of 0.58 was proposed by Melamed (1999) for aligning bitexts for statistical machine translation. The reason for this cutoff is to prevent the LCSR's inherent bias towards shorter words. For example, the word pairs *saw/osa* and *jacinth/hyacinthe*[3] have the same LCSR of $2/3$ and $4/6$ which is counter-intuitive. If the words are identical, then the LCSR for the longer pair and the short pair are the same. A word alignment tool like GIZA++ aligns the words which are *probable translations of each other* in a particular sentence.

Given cognate lists for two languages, the distance between two languages l_a, l_b can be expressed using the following equation:

$$Dist(l_a, l_b) = 1 - \frac{\sum_i sim(l_a^i, l_b^i)}{N} \qquad (1)$$

$sim(l_a^i, l_b^i)$ is the similarity between the ith cognate pair and is in the range of $[0, 1]$. String similarities is only one of the many possible ways for computing the similarity between two words. N is the number of word pairs being compared. Lexicostatistics is a special case of above equation where the range of the sim function is $0|1$. The choice of the similarity function is a tricky one. It would be suitable to select a function which is symmetric. Another criterion that that could be imposed is $sim(x, y) \rightarrow [0, 1]$ where x, y are two strings (or cognate pairs).

To the best of our knowledge, there is no previous work using these lexical similarities for estimating the distances between the languages from a parallel corpus. Section 4 describes the creation of the dataset used in our experiments. Section 5 describes the experiments and the results obtained. Finally the paper concludes with a direction for future work.

4 Dataset

The dataset for these experiments is the publicly available Europarl corpus. The Europarl corpus is a parallel corpus sentence aligned from English to ten languages, Danish, Dutch, Finnish, French, German, Greek, Italian, Portugese, Spanish, and Swedish. Greek was not included in this study since it would have to be transliterated into the Latin alphabet.[4] The corpus was tokenized and the XML tags were removed using a dedicated Perl script. The next task was to create parallel corpora between all the 45 pairs of languages. English was used as the bridge language for this purpose. For each language pair, a sentence pair was included, if and only if there is a English sentence in common to each sentence. Only the first 100,000 sentence pairs for every language pair were included in these experiments. Sentence pairs with a length greater than 40 words were not included in the final set.

All the languages of the Europarl corpus belong to the Indo-European language family, with one exception: Finnish is a member of the Finno-Ugric branch of the Uralic language family, which is not demonstrably related to Indo-European. The other languages in the Europarl corpus fall under three different branches of Indo-European:

1. Danish, Dutch, English, German and Swedish are Germanic languages and can be further subgrouped into North Germanic (or Scandinavian) – Danish and Swedish – and West Germanic – Dutch, English and German, with Dutch and German forming a more closely related subgroup of West Germanic;

[3] Taken from Kondrak (2005)

[4] This is a task for the future.

	pt	it	es	da	nl	fi	fr	de	en
sv	3295	4127	3648	12442	5568	**2624**	3159	3087	5377
pt		10038	13998	2675	2202	**831**	6234	1245	6441
it			11246	3669	3086	**1333**	7692	1738	7647
es				3159	2753	**823**	6933	1361	7588
da					6350	**2149**	3004	3679	5069
nl						**1489**	2665	3968	4783
fi							**955**	**1043**	**1458**
fr								1545	6223
de									2206

sv : Swedish, **pt** : Portugese, **it** : Italian, **es** : Spanish, **da** : Danish, **nl** : Dutch
fi : Finnish, **fr** : French, **de** : German

Table 1: Number of cognate pairs for every language pair.

2. French, Italian, Portuguese and Spanish are Romance languages, with the latter two forming a more closely related Ibero-Romance subgroup, joining French at the next level up in the family tree, and Italian being more distantly related to the other three;

3. Greek forms a branch of its own (but was not included in our experiment; see above).

We would consequently expect our experiments to show evidence of this grouping, including the isolated status of Finnish with respect to the other Europarl corpus languages.

5 Experiments

The freely available statistical machine translation system MOSES (Koehn et al., 2007) was used for aligning the words. The system also extracts the word alignments from the GIZA++ alignments and computes the conditional probabilities for every aligned word pair. For every language pair, the word pairs that have an LCSR value smaller than the *cutoff* are discarded. Table 1 shows the number of pairwise cognates.

We experiment with three string similarity measures in this paper. Levenshtein distance and LCSR are described in the earlier sections. The other measures are *Dice* and *LCSR*. *Dice* is defined as twice the total number of shared character bigrams between two words divided by the total number of bigrams. In the next step, the normalized Levenshtein distance (NLD) between the likely cognate pairs are computed for every language pair. The Levenshtein distance between two words is normalized by the maximum of the length of the two words to account for the length bias. The distance between a language pair is the mean of all the word pairs' distances. The distance results are shown in table 2. *Dice* and *LCSR* are similarity measures and lie in the range of $[0, 1]$.

We use these distances as input to a hierarchical clustering algorithm, UPGMA available in PHYLIP (Felsenstein, 2002), a phylogeny inference package. UPGMA is a hierarchical clustering algorithm which infers a *ultrametric* tree from a distance matrix.

6 Results and discussion

Finnish is clearly the outlier when it comes to shared cognate pairs. This is shown in bold in table 1. Not surprisingly, Finnish shares the highest number of cognates with Swedish, from which it has borrowed extensively over a period of several hundred years. Table 2 shows the pair-wise language distances. The last column shows the language that has the maximum and minimum similarity for each language and distance.

Figures 1, 2 and 3 show the trees inferred on the basis of the three distance measures. Every tree has Spanish, Portegese and Italian under one subgroup, and Danish, Swedish and German are grouped together in all three trees. Finnish is the farthest group in all the trees except in tree 2. The closest languages are Danish and Swedish which are grouped together. Spanish and Portegese are also grouped as close relatives. The trees are not perfect: For instance, French, English and Dutch are grouped together in all the trees.

One can compare the results of these experiments with the tree inferred using Swadesh lists,

	pt	it	es	da	nl	fi	fr	de	en	max	min
sv	0.2994	0.2999	0.306	0.2012	0.2806	0.3131	0.2773	0.2628	0.282	da	fi
	0.5849	0.5876	0.5869	0.6805	0.61	0.6215	0.6187	0.634	0.6195	da	pt
	0.7321	0.7272	0.7264	0.8127	0.7516	0.7152	0.7496	0.7577	0.7424	da	fi
pt		0.2621	0.187	0.2944	0.2823	0.3234	0.2747	0.2783	0.2895	es	fi
		0.6147	0.6824	0.5892	0.6102	0.5709	0.5711	0.5958	0.6008	es	fi
		0.7646	0.8289	0.7289	0.7529	0.7109	0.7541	0.7467	0.7405	es	fi
it			0.2611	0.2923	0.2858	0.3418	0.2903	0.283	0.2802	es	fi
			0.6137	0.5871	0.5916	0.5649	0.5725	0.5847	0.6065	pt	fi
			0.7638	0.7321	0.7474	0.6954	0.7397	0.7448	0.7473	pt	fi
es				0.2965	0.2918	0.3265	0.2725	0.2756	0.2841	it	fi
				0.5924	0.5992	0.5746	0.5799	0.5967	0.6084	pt	fi
				0.7298	0.7444	0.7081	0.7601	0.75	0.7475	pt	fi
da					0.2829	0.3174	0.2596	0.2648	0.269	sv	fi
					0.6064	0.6196	0.6208	0.6164	0.6201	sv	fi
					0.7518	0.7127	0.7639	0.7618	0.7509	sv	fi
nl						0.3343	0.2452	0.2699	0.268	fr	fi
						0.5743	0.6457	0.5971	0.6207	fr	fi
						0.7058	0.7843	0.765	0.7616	fr	fi
fi							0.3369	0.3389	0.3218	sv	it
							0.5525	0.5817	0.6093	sv	fr
							0.7027	0.7135	0.7072	sv	it
fr								0.2734	0.2328	en	fi
								0.5964	0.6505	en	fi
								0.7555	0.7905	en	fi
de									0.2733	sv	fi
									0.6082	sv	fi
									0.749	da	fi

Table 2: The first, second and third entry in each cell correspond to Levenshtein distance, Dice and LCSR distances.

e.g. the results by Dyen et al. (1992), which on the whole agree with the commonly accepted subgrouping of Indo-European (except that according to their results, English is equally far apart from Dutch/German and Danish/Swedish). However, for its successful application to language subgrouping problems, Swadesh lists rely on a large amount of expert manual effort, both in the compilation of a Swadesh list for a new language[5] and in making the cognacy judgements required for the method used by Dyen et al. (1992) and others.

Working with corpora and automated distance measures, we are in a position both to bring more languages into the comparison, and avoiding the admitted subjectivity of Swadesh lists,[6] as well as potentially being able to draw upon both quantitatively and qualitatively richer linguistic data for the purposes of genetic classification of languages.

Instead, we compare our results with the only similar previous work that we are aware of, viz. with the tree obtained by Koehn (2005) from BLEU scores. Koehn's tree gets the two major branches of Indo-European – Germanic and Romance – correct, and places Finnish on its own. The subgroupings of the major branches are erroneous, however: Spanish is grouped with French instead of with Portugese, and English is grouped

[5] It is generally not a straightforward task to determine which item to list for a particular sense in a particular language, whether to list more than one item, etc.

[6] The Swadesh lists were originally compiled on the basis of linguistic experience and intuition about which senses should be universally available as words in languages and which words should be most resistant to replacement over time. These assumptions are only now beginning to be subjected to rigorous empirical testing by typological linguists, and it seems that both may be, if not outright false, then at least too simplistic (Goddard, 2001; Evans and Levinson, 2009; Haspelmath and Tadmor, 2009).

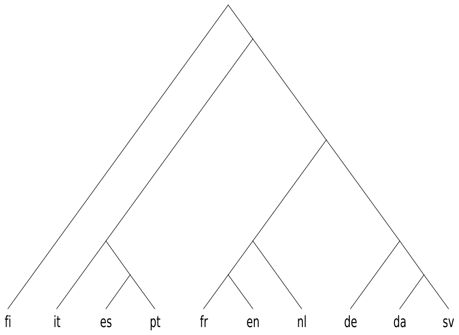

Figure 1: UPGMA clustering for Levenshtein distance scores

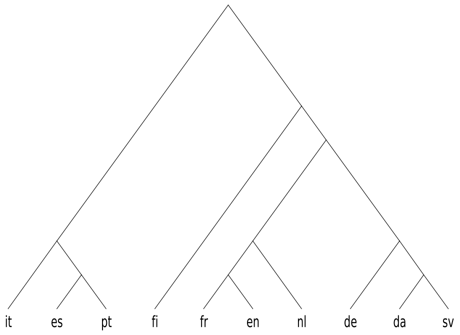

Figure 2: UPGMA clustering for Dice distance scores

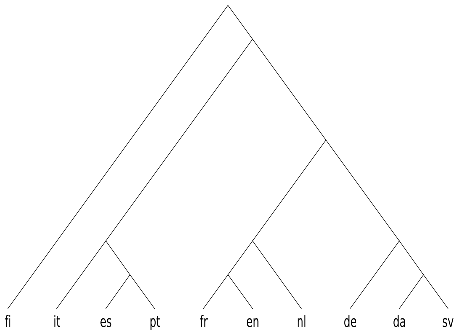

Figure 3: UPGMA clustering for LCSR distance scores

with Swedish and Danish instead of forming a group with German and Dutch.

Using corpora rather than carefully selected word lists brings noise into the comparison, but it also promises to bring a wealth of additional information that we would not have otherwise. Specifically, moving outside the putative core vocabulary, we will pick up evidence of language contact in the form of borrowing of vocabulary and historical spread of orthographical conventions. Thus, one possible explanation for the grouping of Dutch, English and French is that the first two have borrowed large parts of the vocabulary used in the Europarl corpus (administrative and legal terms) from French, and additionally in many cases have a spelling close to the original French form of the words (whereas French loanwords in e.g. Swedish have often been orthographically adapted, for example French *jus* ∼ English *juice* ∼ Swedish *sky* ‘meat juice’).

7 Conclusions and future work

We have presented preliminary experiments with different string similarity measures over translation equivalents automatically extracted from a parallel corpus for estimating the genetic distances among languages. The preliminary results indicate that a parallel corpus could be used for this kind of study, although because of the richer information that a parallel corpus provides, we will need to look into, e.g., how cognates and loanwords could be distinguished. This is an exciting area for future research.

In this study, only the lexical features of the parallel corpora have been exploited, following the tradition of Swadesh list based language comparison. However, using corpora we can move well beyond the lexical level, as corpora can also be used for comparing other linguistic features. Consequently, we plan to experiment with syntactic features such as POS tags for estimating the similarity among languages. Not only the orthographic similarity but also the co-occurrence context vectors for the words could be used to estimate the similarity between translationally similar words.

References

S. Bergsma and G. Kondrak. 2007. Alignment-based discriminative string similarity. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 656.

A. Bouchard-Côté, P. Liang, T.L. Griffiths, and D. Klein. 2007. A probabilistic approach to diachronic phonology. In *Empirical Methods in Natural Language Processing*.

C. Brew, D. McKelvie, et al. 1996. Word-pair extraction for lexicography.

I. Dyen, J.B. Kruskal, and P. Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. American Philosophical Society.

T.M. Ellison and S. Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 273–280. Association for Computational Linguistics.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492.

J. Felsenstein. 2002. PHYLIP (phylogeny inference package) version 3.6 a3. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*.

Cliff Goddard. 2001. Lexico-semantic universals: A critical overview. *Linguistic Typology*, pages 1–65.

R.D. Gray and Q.D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the world's languages: A comparative handbook*. De Gruyter Mouton.

E.W. Holman, S. Wichmann, C.H. Brown, V. Velupillai, A. Müller, and D. Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 9–16. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

G. Kondrak, D. Marcu, and K. Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 46–48. Association for Computational Linguistics.

G. Kondrak. 2002. Algorithms for language reconstruction.

G. Kondrak. 2005. Cognates and word alignment in bitexts. *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 305–312.

A.M.S. McMahon and R. McMahon. 2005. *Language classification by numbers*. Oxford University Press, USA.

I.D. Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):130.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

D. Ringe, T. Warnow, and A. Taylor. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

M. Simard, G.F. Foster, and P. Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pages 1071–1082. IBM Press.

A.K. Singh and H. Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 40–47. Association for Computational Linguistics.

Anil Kumar Singh, Harshit Surana, and Karthik Gali. 2007. More accurate fuzzy text search for languages using abugida scripts. In *Proceedings of ACM SIGIR Workshop on Improving Web Retrieval for Non-English Queries*, Amsterdam, Netherlands.

S. Wichmann, E.W. Holman, D. Bakker, and C.H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*.