An Example Based Machine Translation System Working on Trigrams

Jorge Kinoshita
Escola Politécnica da Universidade de São Paulo
Departamento de Computação e Sistemas Digitais (PCS)
address: Rua Itapeva 38, São Paulo-SP CEP 01332-000, Brazil
email: *jkinoshi@pcs.usp.br*

Abstract

We propose an Example-Based Machine Translation system. The examples were extracted from the Bible, book of Matthew, given in Greek, English and Portuguese and annotated according to Strong's annotation. The examples were arranged as words, bigrams and trigrams. Given a new sentence, we translated n-grams ($1 \leq n \leq 3$) according to the examples and assembled them producing the translated sentence. We tested this procedure by translating the book of Mark from Greek to English and to Portuguese based only on the verses of Matthew. In this paper, we present the chapter 1 of Mark translated by this way. The procedure is very easy to implement and the results show a promising way for translating.

keywords: Example-based machine translation, n-grams, finite state machine.

1 Introduction

The reference (0) presents many methods used in machine translation. In particular, "Example-based machine translation proposes translation by examples collected from past translations." Our system (our research in machine translation started from (0)) uses examples collected from the Bible (0) in Greek and its translations in Portuguese and English. The examples are annotated with numbers (James Strong annotation) which yields a corpus alignment. From the examples we extract words, bigrams and trigrams with their translations and assemble a finite state machine. The translation of a new sentence is done by identifying the words, bigrams and trigrams that were already seen and assigning the corresponding translations. We assemble these translations in order to get the translated sentence.

The reference (0) presents several tasks in which N-grams has been used (in his own words):
1) part of speech tagging
2) finding collocations of technical terms for use in machine translation
3) constraining language models, aiding disambiguation, retrieving texts from large databases, and compiling lexico-syntactic facts for identifying normal and conventional uses of a given word
4) aiding disambiguation of prepositional phrase attachment
5) estimating the probability of previously unseen collocations based on "similar" words.

Although the idea of using n-grams for translating is very simple, so far we did not find a reference to a previous work. We would like to add "translating" to the above list. We choose the Bible as the basis for our examples because it has been studied for centuries, translated to many idioms and in particular, James Strong provided a careful annotation that can be used to make a precise alignment between many versions of the Bible given in many idioms.

2 Strong's annotation

To each word in Greek (or Hebrew), James Strong assigned a number. For example, the original verse in Greek is:
John 1:1 en [1722] arch [746] hn [2258] (5713) o [3588] logoj [3056] kai [2532] o [3588] logoj [3056] hn [2258] (5713) proj [4314] ton [3588] qeon [2316] kai [2532] qeoj [2316] hn [2258] (5713) o [3588] logoj [3056]

which was translated to English as:
John 1:1 ¶ In [1722] the beginning [746] was [2258] (5713) the Word [3056], and [2532] the Word [3056] was [2258] (5713) with [4314] God

[2316], and [2532] the Word [3056] was [2258] (5713) God [2316].

or in Portugese as:
João 1:1 ¶ No [1722] princípio [746] era [2258] (5713) o Verbo [3056], e [2532] o Verbo [3056] estava [2258] (5713) com [4314] Deus [2316], e [2532] o Verbo [3056] era [2258] (5713) Deus [2316].

Looking the numbers, we can identify, for instance, that the greek word "logoj" (number 3056) was translated to "the Word" in English or "o Verbo" in Portuguese. However, we note that "logoj" can also be translated in a different way according to the context. For instance, in Matthew 5:37 "logoj" was translated to "communication" in English or "word" in Portuguese. We present Matthew 5:37 in Greek, English and Portuguese:

Matthew 5:37 estw [2077] (5749) de [1161] o [3588] logoj [3056] umwn [5216] nai [3483] nai [3483] ou [3756] ou [3756] to [3588] de [1161] perisson [4053] toutwn [5130] ek [1537] tou [3588] ponhrou [4190] estin [2076] (5748)

Matthew 5:37 But [1161] let [2077] [0] your [5216] communication [3056] be [2077] (5749), Yea [3483], yea [3483]; Nay [3756], nay [3756]: for [1161] whatever is more [4053] than these [5130] cometh [2076] (5748) of [1537] evil [4190].

Mateus 5:37 Seja [2077] (5749), porém [1161], a tua [5216] palavra [3056]: Sim [3483], sim [3483]; não [3756], não [3756]. O que disto passar [4053] [5130] vem [2076] (5748) do [1537] maligno [4190].

We expect that using bigrams or trigrams we can identify a better context and thus getting a better translation.

3 The trigrams

Given a verse, we break it in trigrams. For instance, given:

John 1:1 en [1722] arch [746] hn [2258] (5713) o [3588] logoj [3056] kai [2532] o [3588] logoj [3056] hn [2258] (5713) proj [4314] ton [3588] qeon [2316] kai [2532] qeoj [2316] hn [2258] (5713) o [3588] logoj [3056]

Table 1: Finite state machine

state	word	next state
0	en	1
1	arch	2
2	hn	3
0	arch	4

Table 2: Translation of the states

state	translation
1	In
2	In - the beginning
3	In - the beginning - was
4	the beginning

we break in the following trigrams:
(en, arch, hn)
(arch, hn, o)
(hn, o, logoj)
and so on.

These trigrams are assembled in a finite state machine. A small fragment of this machine is given in Table 1.

To each state, besides the initial state "0", we assign a translation according to the Strong's annotation. The translation of the previous states is given in Table 2.

It is possible to have multiple translations to the same state, therefore we count the number of times that each translation was used. We have another structure that has the best translation (the translation that appeared most) to each state. A translation assigned to one state may correspond to a translation of a word, bigram or trigram depending on how far the state is from the initial state.

4 The translation

For this experiment we start from the very simple idea that translation is just "exchanging words". We use the finite state machine to identify words, bigrams and trigrams. For each "matching" we assign a translation. Therefore it is possible to have many translations to only one greek word. The preference is for the translation given by a trigram, and then by a bigram and at last by a single word. When 2 trigrams suggests 2 translations for a same word, then we get the suggestion from the most right trigram (and so for bigrams and words).

We are not concerned in exchanging the order of the words in the translated sentence (in Greek, it is normal to put the possessive pronoun after the substantive), although this information can be taken from the examples. A very easy approach is to exchange word order according to the n-grams, however, we loose uniformity. For instance, for some cases the possessive pronoun will appear in the right position and for another cases in the wrong position. We choose to have more uniformity (and simplicity) than some cases in a right way. We are still studying how it can be done and we are presenting new results in a near future.

5 Results

The book of Mark was translated from Greek to Portuguese and to English using our approach. The examples were taken from the book of Matthew in Greek, English and Portuguese. Here, we present the translation of Chapter 1 of Mark and compare both translations.

We present just the translation of Mark, chapter 1, in Portuguese and in English.

In Portuguese:

1:1 é o princípio [euaggeliou] de Jesus Cristo filho de Deus

1:2 como está escrito nos [profhtaij] Eis aí eu envio mensageiro o meu diante face da tua o qual preparará caminho o teu diante de ti

1:3 Voz do que clama no deserto Preparai o caminho do Senhor endireitai veredas as suas

1:4 João [baptizwn] no deserto e pregando [metanoiaj] para remissão de pecados

1:5 e saíam a ter com ele toda a Judéia região e [ierosolumitai] e batizados todos no Jordão [potamw] por ele confessando pecados

1:6 andava porém Ele [endedumenoj] [tricaj] de camelo e um cinto de couro e comia [akridaj] e mel silvestre

1:7 e [ekhrussen] dizendo vem mais poderoso do que eu após as minhas cujas não sou digno [kuyaj] [lusai] [imanta] [upodhmatwn] [autou]

1:8 Eu [ebaptisa] vos com água a sua batizará vos com o Espírito Santo

1:9 que sobreveio dias veio Jesus de Nazaré Galiléia e [ebaptisqh] de João para o Jordão

1:10 e logo para subir da água viu [scizomenouj] [ouranouj] e o Espírito como pomba descendo sobre ele

1:11 e uma voz dos céus tu És Filho o meu amado em quem me comprazo

1:12 e logo o Espírito expele ao deserto

1:13 e ficou lá no deserto dias quarenta [peirazomenoj] Satanás e estava com [qhriwn] e os anjos serviram o

1:14 depois Mas [paradoqhnai] a João Veio Jesus para a Galiléia pregando o evangelho do reino de Deus

1:15 e disse [peplhrwtai] tempo e está próximo o reino de Deus Arrependei-vos e Credes no [euaggeliw]

1:16 Caminhando junto ao mar da Galiléia viu Simão e André que lançavam as redes no mar eram porque pescadores

1:17 e Jesus Vinde após mim e farei eu vos tornar-se pescadores de homens

1:18 e imediatamente eles deixaram as redes o seguiram

1:19 e Passando adiante [oligon] viu Tiago filho de Zebedeu e João aos irmãos as que estavam no barco consertando as redes

1:20 e logo chamou-os deixando -o pai o [zebedaion] que estavam no barco em companhia [misqwtwn] passaram após [autou]

1:21 e [eisporeuontai] em Cafarnaum e logo em dia de sábado entrou na sinagoga [edidasken]

1:22 e se maravilhavam da doutrina sua porque ensinava ele as como autoridade quem tem e não como os escribas

1:23 e ficou na sinagoga pelo Espírito [akaqartw] e [anekraxen]

1:24 dizendo [ea] Que temos nós [nazarhne] Vieste matar conheço te quem És [agioj] de Deus

1:25 E repreendeu Jesus dizendo [fimwqhti] e [exelqe]

1:26 e [sparaxan] o Espírito imundo e [kraxan] voz em alta saiu

1:27 e [eqambhqhsan] todos para [suzhtein] o que significa isto quem doutrina [kainh] que em autoridade e [pneumasin] [akaqartoij] [epitassei] e obedecem lhe

1:28 saiu Portanto fama a sua logo por toda a circunvizinhança da Galiléia

1:29 e logo de [sunagwghj] saindo estando já no outro lado, chegaram na casa de Simão e [andreou] com de Tiago e [iwannou]

1:30 mas [penqera] de Simão [katekeito] [puressousa] e logo Responderam-lhe [authj]

1:31 Então aproximando-se [hgeiren] havendo prendido pela mão e deixou a a febre imediata-

mente passou a servi-lo
1:32 da tarde Ao cair [edu] o sol [eferon] todos e [daimonizomenouj]
1:33 e a cidade todo [episunhgmenh] era que a porta
1:34 curou muitos de várias enfermidades e demônios muitas vezes expeliu e não [hfien] a falar demônios que [hdeisan] [auton]
1:35 e pela manh [ennucon] grandemente Ele se levantou saiu E partiu para deserto um lugar ali [proshuceto]
1:36 e [katediwxan] o Simão companheiros
1:37 e [eurontej] Responderam-lhe Porque todos [zhtousin] [se]
1:38 E disse-lhes vamos pelas [ecomenaj] [kwmopoleij] ali [khruxw] para isto porque [exelhluqa]
1:39 e ficou pregando nas sinagogas por toda a Galiléia e demônios [ekballwn]
1:40 E voltando para que um leproso implorando e que se ajoelhou que se quiseres podes purificar-me
1:41 porém Jesus compadecendo-se estendendo a mão tocou-lhe e disse lhe Quero fica limpo
1:42 e [eipontoj] imediatamente correu da sua lepra e [ekaqarisqh]
1:43 e [embrimhsamenoj] imediatamente expeliu [auton]
1:44 e disse lhe Olha a ninguém Não não o digas mas vai mostrar-te ao sacerdote e fazer a respeito [kaqarismou] o teu o que ordenou que Moisés para servir de testemunho ao povo
1:45 saindo começou a pregar muitas vezes e [diafhmizein] a palavra para [dunasqai] [fanerwj] em cidade entrares mas fora no [erhmoij] [topoij] era e [hrconto] [pantacoqen]

in English:
1:1 are the beginning [euaggeliou] of Jesus Christ the son of God
1:2 as it is written in [profhtaij] Behold I send messenger my before face thy who shall prepare way thy before thee
1:3 The voice of one crying in the wilderness Prepare ye the way of the Lord straight make paths his
1:4 it came to pass John [baptizwn] in the wilderness and preaching with the baptism [metanoiaj] for the remission of sins
1:5 and went out to him all Judaea the region and [ierosolumitai] And were baptized all in Jordan [potamw] by him confessing sins their
1:6 there was But John [endedumenoj] [tricaj] camel's and belt a leather about loins his and eating [akridaj] and honey wild
1:7 and [ekhrussen] saying then cometh mightier than I after my whose not I am worthy [kuyaj] [lusai] [imanta] [upodhmatwn] [autou]
1:8 I indeed [ebaptisa] you in water the same And shall baptize you in Spirit the Holy
1:9 And there arose in those days came Jesus of Nazareth Galilee and [ebaptisqh] by John into Jordan
1:10 And immediately going up out of the water he saw [scizomenouj] [ouranouj] and the Spirit like a dove descending upon him
1:11 and a voice it came to pass from heaven thou Art Son my beloved in whom I am well pleased
1:12 and immediately the Spirit him bringeth forth into the wilderness
1:13 And was there in the wilderness days forty [peirazomenoj] by Satan And man was with [qhriwn] and the angels ministered to him
1:14 after But [paradoqhnai] John came when Jesus into Galilee preaching the gospel of the kingdom of God
1:15 And saying [peplhrwtai] the time and is at hand the kingdom of God Repent ye and Believe ye in [euaggeliw]
1:16 walking And by the sea of Galilee saw Simon and Andrew brother his casting a net in the sea they were for fishermen
1:17 And said to them Jesus Follow me and I will make you be fishers of men
1:18 And they immediately left nets their and followed him
1:19 And going on from there [oligon] he saw James the son of Zebedee and John brother his and them in a boat mending nets
1:20 And immediately he called them and left father their [zebedaion] in a boat with [misqwtwn] they went after [autou]
1:21 and [eisporeuontai] in Capernaum And immediately on the sabbath he went in into synagogue [edidasken]
1:22 and this, they were astonished at doctrine his For he taught them as authority one having and not as the scribes
1:23 And was in synagogue their man in Spirit [akaqartw] and [anekraxen]
1:24 saying [ea] What have we to do with thee

Jesus [nazarhne] art thou come to destroy us I know if [agioj] of God
1:25 And rebuked to him Jesus saying [fimwqhti] and [exelqe] of it
1:26 and [sparaxan] him the Spirit the unclean and [kraxan] voice with a loud came out of it
1:27 and [eqambhqhsan] all so that [suzhtein] to them saying what that meaneth this doctrine [kainh] that in power and [pneumasin] [akaqartoij] [epitassei] and obey him
1:28 came out Therefore fame his immediately into all surrounding country of Galilee
1:29 And immediately of [sunagwghj] went out they came into the house of Simon and [andreou] with of James and [iwannou]
1:30 but [penqera] of Simon [katekeito] [puressousa] And immediately They say to him concerning [authj]
1:31 And came [hgeiren] her had laid hold hand her and left her the fever immediately and ministered to them
1:32 the evening When was come when [edu] the sun [eferon] to him all sick people and [daimonizomenouj]
1:33 and city all [episunhgmenh] was to who door
1:34 and he healed many sick people various diseases and demons many things he cast out and not [hfien] spoke demons that [hdeisan] [auton]
1:35 And in the morning [ennucon] was exceeding he arose came out And and departed into a desert place and there [proshuceto]
1:36 and [katediwxan] him Simon and they that were with him
1:37 and [eurontej] him They say to him Because all [zhtousin] [se]
1:38 And he saith to them let us be going into [ecomenaj] [kwmopoleij] that and there [khruxw] into this For [exelhluqa]
1:39 And was preaching in synagogues their into all Galilee and demons [ekballwn]
1:40 And he cometh to him a leper beseeching him and kneeling down him and saying to him That if thou wilt thou canst me clean
1:41 But Jesus was moved with compassion put forth his hand and touched him And saith to him I will be thou clean
1:42 and [eipontoj] his immediately spread from his leprosy and [ekaqarisqh]
1:43 and [embrimhsamenoj] to him immediately he cast out [auton]
1:44 And saith to him See no man nothing thou tell but go thyself show to the priest and offer for [kaqarismou] thy commanded Moses for a testimony to them
1:45 But he went out began to preach many things and [diafhmizein] the word so that henceforth him [dunasqai] [fanerwj] in a city enter but outside in [erhmoij] [topoij] was and [hrconto] to him [pantacoqen]

6 Conclusion

Although there are many errors in the translation, it is possible to understand many verses. The results achieved this quality because the examples are in the same domain (Matthew and Mark are very close). Strong's annotation was used just for the alignment of the corpus. The same procedure can be applied using any other way of alignment ((0), (0)). Strong's annotation provides a richer information that we are not using: his annotation refers to the stems, so for instance, the word "komoj", "kosmw", "kosmon" and "kosmou" are assigned to the same number ([2889]). This information can be used in future improvements of our system. The Portuguese translation shows to be better than the English translation due to:
- English word order are much more severe than in Portuguese.
- The correspondence between words are more straight in Greek-Portuguese than Greek-English. For instance: English verbs need the subject in a explicit way causing the insertion of pronouns in the text. In Matthew 2:9 the greek word "eporeuqhsan" [4198] (5675) was translated as "partiram" in Portuguese and as "they departed" in English. Another difference is the use of modals. Matthew 2:9 is presented in Greek, English and Portuguese:

Matthew 2:9 oi [3588] de [1161] akousantej [191] (5660) tou [3588] basilewj [935] eporeuqhsan [4198] (5675) kai [2532] idou [2400] (5628) o [3588] asthr [792] on [3739] eidon [1492] (5627) en [1722] th [3588] anatolh [395] prohgen [4254] (5707) autouj [846] ewj [2193] elqwn [2064] (5631) esth [2476] (5627) epanw [1883] ou [3757] hn [2258] (5713) to

[3588] paidion [3813]

Matthew 2:9 ¶ When [1161] they had heard [191] (5660) the king [935], they departed [4198] (5675); and [2532], lo [2400] (5628), the star [792], which [3739] they saw [1492] (5627) in [1722] the east [395], went before [4254] (5707) them [846], till [2193] it came [2064] (5631) and stood [2476] (5627) over [1883] where [3757] the young child [3813] was [2258] (5713).

Mateus 2:9 ¶ Depois [1161] de ouvirem [191] (5660) o rei [935], partiram [4198] (5675); e [2532] eis [2400] (5628) que a estrela [792] que [3739] viram [1492] (5627) no [1722] Oriente [395] os [846] precedia [4254] (5707), até que [2193], chegando [2064] (5631), parou [2476] (5627) sobre [1883] onde [3757] estava [2258] (5713) o menino [3813].

For our approach the closer the 2 languages, the better. Hence, this kind of translation must be very suitable for Portuguese-Spanish texts.

References

The EAGLES Lexicon Interest Group. (no date). Machine Translation [Online]. Available: http://www.ilc.pi.cnr.it /EAGLES96/rep2/node31.html [1999, April 16].

Lynellen D.S.P. Smith. (Summer 1994). Estimation of N-gram Probabilities [Online]. Available: http://www2.msstate.edu /~lds3/write/ngrams.html [1999, April 16].

KINOSHITA, J.: Aspectos de Implementação de Uma Ferramenta de Auxílio à Tradução Inglês-Português. Tese de Doutorado, Departmento de Computação e Sistemas Digitais da Escola Politécnica da Univesidade de São Paulo, Brazil (1997)

Dan Melamed, A Geometric Approach to Mapping Bitext Correspondence, IRCS Technical Report #96-22, a revised version of the paper presented at the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA, May 1996.

Dan Melamed, Empirical Methods for MT Lexicon Construction, in L. Gerber and D. Farwell, Eds., Machine Translation and the Information Soup, Springer-Verlag, 1998.

Bíblia Online. CD-ROM - versão 1.0. Sociedade Bíblica do Brasil. ISBN 1-896755-04-6 (1997)