Sentiment analysis for reviews and microtexts based on lexico-syntactic knowledge

David Vilares
Departamento de Computación, Universidade da Coruña
Campus de Elviña, 15011, A Coruña
Spain
david.vilares@udc.es

We describe two methods to perform sentiment analysis both on long and short texts written in Spanish language. We first present an unsupervised method based on dependency parsing which calculates the semantic orientation (SO) of the sentences in order to classify the polarity. We then propose a hybrid approach which uses the computed SO and lexico-syntactic knowledge as features for a supervised classifier. Experimental results show the utility of employing syntactic information to classify the polarity in both types of texts and the importance of defining mechanisms to adapt the system for a specific domain and social medium.

Sentiment Analysis, Opinion Mining, Dependency Parsing, Machine Learning

1. INTRODUCTION

With the apparition of Web 2.0 and the rise of blogs, forums and social networks, users express their views about various topics on these sites. They discuss current issues and praise, compare or complain about products, services and even people. The economic benefits that can be derived from this knowledge are obvious, so the market has begun to demand solutions to analyse this enormous flow of opinions. In this respect, *sentiment analysis* (SA) is a growing field of research focussed on automatic processing of subjective information, where one of the main tasks is *polarity* classification, *i.e.*, to determine whether the opinion expressed is positive, negative, neutral or mixed. There is no standardisation about the polarity categories, but most of studies perform a binary (positive, negative) or a ternary classification (positive, negative, neutral), although there is also related research which takes into account more categories.

The polarity classification task has been tackled in the last decade from two different perspectives: supervised machine learning (ML) approaches and non-supervised semantic-based methods. ML solutions involve building classifiers from a collection of annotated texts (Pang et al. (2002)), where each text is usually represented as a bag-of-words. It is also common to include some linguistic-related processing for preparing features (Bakliwal et al. (2012)). The main drawback of this angle is that it is highly domain dependent (Taboada et al. (2011)). On the other side, semantic-based approaches (Turney (2002)) involve the use of dictionaries where different kinds of words are tagged with their semantic orientation (SO); they have been applied successfully in many contexts but their performance is not optimum because different application domains and social media have many specific subjective elements, this results in a low recall of the opinion lexicons (Zhang et al. (2011)).

Traditionally, SA research has focussed on long texts. For example, Taboada et al. (2011) propose a lexicon-based method which deals with phenomena such as intensification, negation or irrealis. With a similar aim, Abbasi et al. (2008) describe an approach which takes stylistic and syntactic components as features for a supervised classifier. However, the recent success of microblogging social networks, such as Twitter, has increased interest in monitoring short texts. In this line, Bakliwal et al. (2012) performed an sentiment scoring algorithm which uses prior information to classify the polarity of tweets, and Sidorov et al. (2012) explore different settings of parameters for a supervised classifier.

In this context, this paper proposes an unsupervised and a supervised approach which are able to perform binary polarity classification over reviews and short texts. We adopt in both cases an NLP perspective which takes into account lexical information and syntactic relations between words. The unsupervised approach is able to treat relevant linguistic phenomena, such as intensification, subordinate adversative clauses or negation, to then calculate the SO of the text. The ML approach uses lexico-syntactic knowledge and the information provided by our unsupervised system as features for a classifier.

The methods proposed in this article have been tested with the following corpora:

- *HOpinion*[1]: A collection of 17,934 hotel reviews, rated between one and five stars. There are 841 *one-star*, 1269 *two-star*, 3468 *three-star*, 6244 *four-star* and 6112 *five-star* reviews. Reviews ranked with one or two stars are considered negative. We discard three-star reviews because they are considered as neutral or mixed reviews. This is a widely accepted strategy that has been employed in other binary polarity classification studies and corpora, like the SFU Spanish Review Corpus[2] (Brooke et al. (2009)). Documents ranked with four or five stars are taken as positive. We employed the 80% of the corpus as the training set and the remaining 20% as the test set.
- *TASS 2012*: This corpus was presented at the Workshop on Sentiment Analysis at SEPLN (Villena-Román et al. (2013)). It is a collection of Spanish tweets written by public figures that is composed of a training and a test set which contain 7,219 and 60,798 tweets, respectively. Each one is annotated with one of these six categories: *strongly positive* (P+), *positive* (P), *neutral* (NEU), *negative* (N), *strongly negative* (N+) or *without opinion* (NONE). In order to homogenise experimental results with HOpinion, we only take into account two polarities: positive (P+, P) and negative (N+, N), discarding the rest of the tweets.

2. NATURAL LANGUAGE PROCESSING TASKS

In order to employ linguistic knowledge in SA, we first need to apply natural language processing (NLP) to the texts. As a previous step, all reviews were pre-processed as follows:

- *Unification of compound expressions*. There are many compound expressions in Spanish like *'sin embargo'* ('however') or *'en absoluto'* ('not at all'), that must usually be interpreted as single units of meaning. To find them, we use a dictionary of compound expressions, extracted from the Ancora corpus (Taulé et al. (2008)). If the pre-processing algorithm identifies a group of these words, it unifies them into a single token (*'en absoluto'* becomes *'en_absoluto'*).
- *Normalization of punctuation marks*. People do not usually respect punctuation rules in web reviews. This is a handicap for the rest of processing. To resolve this, pre-processing homogenises all punctuation mark representation by adding blanks when required.
- *Emoticon replacement*: We employ the emoticon collection published in (Agarwal et al. (2011)). Each emoticon is replaced with one of these five labels: strong positive (ESP), positive (EP), neutral (ENEU), negative (EN) or strong negative (ESN).
- *Most frequent unrecognised abbreviations spell-checking*: We replace some of the most habitual ungrammatical Spanish abbreviations by their grammatical form. For example, *'q'* becomes *'que'* ('that'), *'pq'* becomes *'porque'* ('because'), ...
- *URL normalisation*: Web addresses are replaced with the string *'URL'*.
- *Laughs normalisation*: Different variants of laughs in Spanish language (*e.g. 'jjjaja'*, 'JJEEJJ',...) are normalised as *jxjx* where x $\in \{a, e, i, o, u\}$.

In addition, we consider an *ad-hoc* pre-processing for treatment of tweets:

- *Twitter usernames ('@')*: User mentions are modified: we eliminate the '@' symbol and capitalise the first character (*e.g. '@user' becomes 'User'*).
- *Hashtags ('#')*: If it appears at the beginning or the end of a tweet, then the complete hashtag is eliminated. Otherwise we only delete the '#'.

As a second step, Part-of-Speech (PoS) tagging is performed by running the Brill tagger (Brill (1992)), using Ancora as the training corpus. A challenge for Spanish PoS tagging is that the use of accents is commonly ignored by people when writing in web reviews. The drawback is that taggers trained with regular corpora are not able to tag pairs of words that should use diacritical accents in order to difference their meaning. To improve performance, we have expanded the training set: we cloned each sentence to obtain its equivalent without any acute accent.

Once these steps have been performed, we use dependency parsing (Gómez-Rodríguez et al.

[1] http://clic.ub.edu/corpus/hopinion

[2] This issue is detailed on the readme file of www.sfu.ca/∼mtaboada/download/downloadCorpusSpa.html

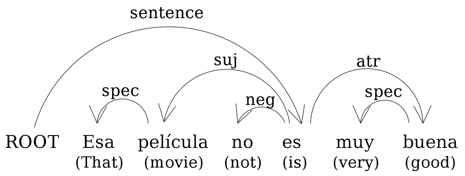

***Figure 1:** Dependency parsing for a Spanish sentence*

(2011)) for analysing the syntactic structure of each given sentence. In particular, we have used MaltParser and the Ancora corpus to train a dependency parser based on the *Nivre arc-eager* algorithm (Nivre (2008)). As a result, we obtain a *dependency tree* for each sentence, consisting of a set of *head/dependent* binary relations, called *dependencies*, between words. Each dependency has a label with a given *dependency type*, which denotes the existing syntactic relation between head and dependent. To simplify computational implementation, an artificial ROOT node is added as the first word of each sentence. Figure 1 shows an example of this type of analysis for the sentence: *'Esa película no es muy buena'* which translates to *'That movie is not too good'*.

3. AN UNSUPERVISED SYSTEM BASED ON THE SEMANTIC ORIENTATION OF THE SENTENCES (US)

Most unsupervised SA systems are typically lexicon-based solutions that cannot interpret the syntactic structure of texts. In order to try to overcome these limitations, it is common to implement heuristics to simulate a comprehension of negation, intensification and other linguistic constructions, but these often fail, given the complexity of human language. As an alternative, in this section we propose an unsupervised, dependency parsing based method for determining the SO of reviews. We use the SODictionariesV1.11Spa (Brooke et al. (2009)) as our opinion lexicon. It is a collection of subjective words where each one has associated an SO between +5 and -5, according to its generic perception (*e.g. 'happiness'* has an SO of +5, *'killer'* has -5 and *'good'* is associated with a value of +2).

3.1. Treatment of intensification

An intensifier is a word or an expression which plays the role of a valence shifter in a sentence. There are two types according to their category: *amplifiers* and *downtoners*. The former maximize SO of one or more tokens, such as *'muy'* ('very'); whereas the latter decrement it, *'en absoluto'* ('not at all') or *'poco'* ('little'). The SODictionariesV1.11Spa have a specific dictionaries for intensifiers, where each intensifier has an associated percentage, positive if it is an amplifier and negative if it is a downtoner. We use syntactic dependencies to identify the scope of an intensifier; whenever an adverb is a dependent of a specifier (*spec, espec*) or an adjunct (*cc, sadv*) type, we take that word as a valence shifter and its head as the exact scope to be shifted. For example, in the Figure 1, the term *'muy'* would modify the SO of *'buena'* by +25%, according to its value in SODictionariesV1.11Spa.

3.2. Treatment of subordinate adversative clauses

A subordinate adversative clause expresses an event or fact that is the opposite to that of the main clause. In an SA context, we hypothesise that these type of constructions are a way of restricting, excluding or amplifying the sentiment reflected by both the main and subordinate clauses. We consider subordinate adversative clauses as a special case of intensification, but involving clauses, not individual terms. In this respect, we distinguish two different types of adversative conjunctions, as is pointed out in Campos (1993), Chapter 3. The first type, *restrictives* (*'but', 'while', . . .*), increase the sentiment of the subordinate clause and decrease the SO of the main clause. The second type, *exclusives* (*'but rather', 'but in the other hand', . . .*), ignore totally the sentiment reflected in the main clause. In this way, our approach is able to calculate coherently the sentiment of sentences such as *'The actor acted badly but the movie was great'*, where the sentiment of *'The actor acted badly'* is partially diminished by the subordinate adversative clause. Modifier percentage of restrictive conjunctions was established by an empirical process over the SFU Spanish Review Corpus. The SO of the main clause is increased by 40% and the SO of the subordinate sentence is decreased by 25%.

3.3. Treatment of negation

The most common and simple way to negate a sequence of tokens in Spanish is the adverb *'no'* ('no'/'not'), but other terms such as *'sin'* ('without') or *'nunca'* ('never') are frequently employed. However, some types of Spanish sentences usually require the use of double negatives to make a negative sentence. In this respect, words like *'nada'* ('nothing'), *'ninguno'* ('none') or *'nadie'* ('nobody') are commonly preceded by *'no'*. Moreover, the difference between a negation term and a downtoner is diffuse. Tokens like *'apenas'* ('barely') or *'casi'* ('almost') could easily be classified in either of these two categories. We have chosen to consider these type of expressions as intensifiers and therefore we only consider explicitly as negators the adverbs *'no'*, *'nunca'* and *'sin'*, which cover a great number of negative sentences. Our treatment of a negation consists of two basic steps: 1) *identify the scope of a* *negation term* and 2) *modify the semantic orientation of affected tokens*.

3.3.1. Scope identification

The procedure for identifying the scope of a negation depends on the adverb used in the phrase. The syntactic structure used in Ancora for representing an adverb *'sin'* assures us that its child node should be the scope of negation, without needing to analyse the dependency type. But we cannot assume the same for the negators *'no'* and *'nunca'*. Normally they are represented as leaf nodes and the candidate scope of negation always involves a head node or a collection of sibling nodes, so we require a more complex algorithm for their treatment. We follow a procedure based on Jia et al. (2009), but we have adapted this procedure to profit from the additional information provided by the syntactic structure of the sentence. We use dependency types to directly extract the scope of negation. When a token has a negator *'no'* ('not') or *'nunca'* ('never') as a child node and it is a dependency of type *'neg'* or *'mod'*; we try a collection of syntactic heuristic rules in the following order:[3]

1. *Subjective parent rule*: Whenever a parent node of a negation term has sentiment, only that node is negated. For example, in the sentence *'he does not praise my work'*, the negation *'not'* depends on *'praise'*, which is included as a subjective word in the SO dictionaries, so we consider this term as the scope of the negation.

2. *Subject complement/Direct object rule*: Whenever a branch at the same level as a negation node is labelled with a dependency of type subject complement (*atr*) (*'the meal is not good'*) or a direct object (*cd*) (*'the meal does not look good'*), our sentiment analyser negates that branch.

3. *Adjunct rule*: Whenever a negation term has an adjunct branch (*cc*) at the same level, the sentiment of that branch is shifted. If there is more than one adjunct, only the first one is negated. For example, in the sentence *'he does not work efficiently on Fridays'*, our method takes the mood adjunct (*'efficiently'*) as the scope of the negation, because it is the nearest to the negation.

4. *Default rule*: If none of the previous rules matches, we consider as scope the sibling branches of a negator.

3.3.2. Polarity flip

We follow a shift negation algorithm where the SO value is shifted toward the opposite polarity by a fixed amount: following Taboada et al. (2011), we have chosen a shift value of 4 for the adverbs *'no'* ('not') and *'nunca'* ('never'). For the adverb *'sin'* ('without'), based on our experimental setup, we have chosen a value of 3.5. We hypothesise this kind of negation as being less potent, given that its scope is fairly local. Experimental results showed an slightly improvement in accuracy when carrying out this strategy.

4. A SUPERVISED SYSTEM BASED ON LINGUISTIC FEATURES

We now propose a hybrid system which combines lexical, syntactic and semantic knowledge with ML techniques. In particular, linguistic features are used to feed an SMO, an implementation of SVM, presented in (Platt (1999)), and incorporated by default in the WEKA[4] data mining software.

4.1. Base supervised system (BSS)

We include the SO obtained by our unsupervised system, and the number of positive and negative words in a text, as features for a supervised classifier. We use the SODictionariesV1.11Spa to determine which words are opinionated.

4.2. Lexico-syntactic features (LSF)

The employment of POS-tagging information in polarity classification tasks is a widely discussed issue. Pak and Paroubek (2010) suggest that certain POS-tags, such as adjectives or personal pronouns, are more frequent in subjective texts. In this respect, we observed a similar tendency in the training sets employed in this paper. Table 1 shows a selection of relevant tag frequencies. In the same way, we hypothesise that dependency types are also useful in order to classify the polarity of the tweets. Table 2 shows the frequency of some dependency types[5] on the HOpinion and TASS 2012 training sets.

Tag	$\mathbf{P}_{HOpinion}$	$\mathbf{N}_{HOpinion}$	$\mathbf{P}_{TASS}$	$\mathbf{N}_{TASS}$
a	0.086	0.066	0.058	0.054
n	0.210	0.195	0.260	0.264
v	0.123	0.142	0.114	0.126
r	0.066	0.068	0.043	0.042

Table 1: *Tag frequencies in the training set: adjectives (*a*), nouns (*n*), verbs (*v*) and adverbs (*r*)*

4.3. Specific domain features (SDF)

Each domain and social medium have some specific elements that denote (implicitly or explicitly)

[3] Only the first matching rule is applied.

[4] http://www.cs.waikato.ac.nz/ml/weka/

[5] We use the Ancora dependency type tags.

Tag	P$_{HOpinion}$	N$_{HOpinion}$	P$_{TASS}$	N$_{TASS}$
atr	0.028	0.125	0.0105	0.000
adjunct	0.071	0.079	0.0495	0.004
neg	0.002	0.003	0.001	0.003
cag	0.001	0.0546	0.001	0.0724

Table 2: *Dependency type frequencies in the training set: subject complement (*atr*), adjunct (*cc*), negation (*neg*) and agent (*cag*)*

sentiment. For example, in tweets, there is a high frequency of some special subjective elements: emoticons, laughs and some Twitter tags, such as *Follow Friday* (*'FF'*) or *Retweet* (*'RT'*), are some of the clearest examples. In the same line, some words are only opinionated in some domains, such as *'air conditioning'*, that it would normally be an objective word, but it is not strange that it could be a polar one in a hotel review (*e.g. 'The room didn't have air conditioning'*). To treat this issue, we have developed an automatic mechanism that enriches and adapts semantic knowledge to a particular field. The goal is to create a ranked list of words to help distinguish between the different polarities, and use each word of that list as a feature for the classifier. We use binary occurrence as the weighting factor in case of tweets, because we hypothesise that each word usually appears at most once in a tweet; and the total occurrence in case of long texts.

Term	Rank$_{TASS}$	Rank$_{HOpinion}$
EP (emoticon)	1	608
FF	30	-
jaja (laugh)	46	35,325
clean	8,997	68
air conditioning	-	78

Table 3: *Ranking of some of discriminating terms on the training set of the* TASS 2012 *corpus*

We rank the terms by measuring the *information gain* with respect to the class, employing the attribute selection tools provided by WEKA and the respective training set. We extracted around 14,000 discriminating terms for the TASS 2012 corpus and about 40,000 in case of HOpinion. However, we saw in both cases that only few hundred of terms were needed for achieving the best performance. Table 3 compares the rank, between HOpinion and the TASS 2012 corpus, for some discriminating terms.

5. EXPERIMENTAL RESULTS

Tables 4 and 5 show the performance on HOpinion and the TASS 2012 test sets.[6] In both corpora, the unsupervised system (US) obtains a good accuracy, but it has a lower performance for negative texts, specially on the HOpinion corpus. This tendency to favour positive classifications is an issue widely discussed on the literature (Brooke et al. (2009)). The employment of the SO and the total number of positive and negative words (+BSS) has a satisfactory effect on the performance, which suggests that the information provided by our unsupervised approach is useful for a supervised classifier. The incorporation of PoS-tag and syntactic information (+LSF) improves the classification performance on positive and negative texts. This reinforces the idea that certain POS-tags and syntactic functions more frequently depending on the polarity of the review. The accuracy obtained by our final approach (+SDF) suggests that, although generic opinion lexicons and the morphosyntactic structure are helpful to classify the sentiment, we need to incorporate domain semantic knowledge to optimise the performance. Moreover, this final version is able to partially counteract the favourable tendency to positive classifications present in the rest of versions. Finally, we compare our methods with a pure ML approach. We trained an SMO (keeping the WEKA default configuration) which takes as features the bag of words of a text. They are pre-processed as indicated in Section 2 and lemmatised, to then use binary occurrence as weighting factor for tweets, and total occurrence for reviews.

Measure	US	+BSS	+LSF	+SDF	SMO
F$_p$	0.893	0.942	0.946	**0.964**	0.961
F$_n$	0.561	0.511	0.549	**0.761**	0.731
Accuracy	0.828	0.897	0.903	**0.938**	0.917

Table 4: *Results on the test set of the HOpinion 2012*

Measure	US	+BSS	+LSF	+SDF	SMO
F$_p$	0.730	0.798	0.825	**0.877**	0.849
F$_n$	0.689	0.671	0.753	**0.833**	0.788
Accuracy	0.711	0.750	0.795	**0.857**	0.824

Table 5: *Results on the test set of the* TASS 2012

6. CONCLUSIONS AND FUTURE WORK

We describe an unsupervised and a hybrid method based on linguistic knowledge. Experimental results suggest that both approaches satisfactorily perform sentiment analysis on reviews and microtexts, which reinforces the utility of employing lexico-syntactic knowledge in order to classify the polarity of opinions.

[6] We used the F-measure defined as $F = \frac{2 \times R \times P}{R+P}$, where P is the number of true positives divided by the sum of true and false positives, and R is the number of true positives divided by the sum of the true positives and false negatives. F_p and F_n refers to F-measure for positive and negative opinions, respectively.

As future work, we would like to incorporate more linguistic phenomena in our unsupervised method, such as the irrealis or the subjunctive mood, as other systems do (Taboada et al. (2011)). In the same line, we think that expanding our treatment of negation, including more negation terms, would have a positive effect. With regard to the supervised system, we desire to explore more thoroughly the employment of syntactic knowledge as features for the classifier.

ACKNOWLEDGMENTS

Research reported in this paper has been partially funded by Ministerio de Economía y Competitividad and FEDER (TIN2010-18552-C03-02) and by Xunta de Galicia (CN2012/008, CN2012/319).

REFERENCES

Abbasi, A., H. Chen, and A. Salem (2008, June). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst. 26*(3), 12:1–12:34.

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. Passonneau (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, Stroudsburg, PA, USA, pp. 30–38. ACL.

Bakliwal, A., P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma (2012). Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Jeju, Korea, pp. 11–18. ACL.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, HLT'91, Stroudsburg, PA, USA, pp. 112–116. Association for Computational Linguistics.

Brooke, J., M. Tofiloski, and M. Taboada (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the International Conference RANLP-2009*, Borovets, Bulgaria, pp. 50–54. ACL.

Campos, H. (1993). *De la oración simple a la oración compuesta: Curso Superior de Gramática Española*. Georgetown University Press.

Gómez-Rodríguez, C., J. Carroll, and D. Weir (2011, September). Dependency parsing schemata and mildly non-projective dependency parsing. *Computational Linguistics 37*(3), 541–586.

Jia, L., C. Yu, and W. Meng (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM'09, New York, NY, USA, pp. 1827–1830. ACM.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Computational Linguistics 34*(4), 513–553.

Pak, A. and P. Paroubek (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pp. 79–86.

Platt, J. C. (1999). Advances in kernel methods. Chapter Fast training of support vector machines using sequential minimal optimization, pp. 185–208. Cambridge, MA, USA: MIT Press.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon (2012). Empirical study of opinion mining in Spanish tweets. In *LNAI 7629-7630*.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics 37*(2), 267–307.

Taulé, M., M. A. Martí, and M. Recasens (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424.

Villena-Román, J., S. Lana-Serrano, J. C. González Cristóbal, and E. Martínez-Cámara (2013). TASS - Worshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural 50*, 37–44.

Zhang, L., R. Ghosh, M. Dekhil, M. Hsu, and B. Liu (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89, HP Laboratories, Palo Alto, CA.